

Moral Contagion: the Influence of Group Behaviour on Individual Honesty

by

Maria Leonor Dinis Neto

i6144836

A Thesis submitted to the
Faculty of Psychology and Neuroscience, Maastricht University

In fulfilment of the requirements for the degree
Research Master in Cognitive and Clinical Neuroscience
Neuroeconomics specialisation

Supervised by

Jean-Claude Dreher, PhD

Institut des Sciences Cognitives Marc Jeannerod, CNRS

Arno Riedl, PhD

School of Business and Economics, Maastricht University

Julien Benistant, PhD

Institut des Sciences Cognitives Marc Jeannerod, CNRS

Institut des Sciences Cognitives Marc Jeannerod

Lyon, France

12 200 words

August 2020

Table of Contents

List of abbreviations	2
Abstract	3
1. Introduction	4
2. Methods	14
2.1. Participants	14
2.2. Trials	14
2.3. Task	16
2.4. Group behaviour manipulation	18
2.5. Analyses	19
2.5.1. Decision model and Basic Model	20
2.5.2. Variations of the Basic Model	20
3. Results	22
3.1. Overall lying	22
3.2. Learning	23
3.3. Contagion	26
3.4. Model-based analysis	29
4. Discussion	35
5. References	41
Appendix A – Additional Figures and Tables	51
Appendix B	55
1. Instructions	55
2. Comprehension questionnaire	58
Appendix C – Details on modelling	60
1. Static Models	61
1.1. Basic Model	61
1.2. Fixed Cost Model	61
1.3. Conformity Model	61
2. Reinforcement Learning Models	62
2.1. Moral- α Model	62
2.2. Moral- δ Model	62
2.3. Moral- α - δ	63
2.4. Conformity-Action	63
2.5. Conformity-Outcome	64

List of abbreviations

ACC – Anterior cingulate cortex

dIPFC – Dorsolateral prefrontal cortex

dmPFC – Dorsomedial prefrontal cortex

fMRI – Functional magnetic resonance imaging

M – Mean

Mdn – Median

mPFC – Medial prefrontal cortex

OR – Odds ratio

PCC – Posterior cingulate cortex

RL – Reinforcement learning

SD – Standard deviation

tDCS – Transcranial direct current stimulation

TPJ – Temporoparietal junction

vlPFC – Ventrolateral prefrontal cortex

vmPFC – Ventromedial prefrontal cortex

Abstract

Moral principles are a structural and guiding aspect of every society and provide common expectations for how to behave. Being honest is one of these principles. As social beings, we are continuously exposed to others' moral and immoral behaviour and learn from it, internalising what is wrong and right, but also what is permissible. Here, we delve into how social norms, learned by observing group behaviour, influence one's own honesty and moral preferences. For this purpose, we experimentally manipulated social norms by presenting individuals with the decisions of two different groups – one honest and one dishonest – and instructing them to predict their behaviour. We recruited 82 participants to take part in an online experiment, consisting of a 2-forced choice honesty paradigm. Participants saw a 6-face die roll and were instructed to report the number seen from two alternatives: the true answer and a more profitable one, as each outcome was associated with a payoff. They first performed the task alone, followed by two blocks in which they also learned about the behaviour of an honest and dishonest group by predicting it. We found an overall low probability of lying, but a high degree of heterogeneity in behaviour. Individuals were very good at predicting others' decisions and were slightly influenced by others' dishonesty, as reflected in higher probability of lying. Using computational modelling, we found that the best way to describe contagion was by introducing a preference bias towards or against lying in the face of social information, on top of one's own moral preferences.

1. Introduction

Discerning between what is right and what is wrong is one of the fundamental concerns of human society and an ancient question that has kindled a branch of philosophy, that is, ethics. Spanning across centuries and schools of thought (e.g., Aristotelian ethics, Ross & Brown, 2009; Kant, 2006; Sen, 2012; Smith, 2009), morality is as relevant as ever to fathom human behaviour and decision-making, as it provides individuals with a set of common expectations, thus facilitating socialisation (FeldmanHall, Son, & Heffner, 2018). A big portion of the work in ethics relates to normative or prescriptive ethics, i.e., how individuals should behave in a particular situation. Psychologists, economists and, more recently, neuroeconomists endorse a more descriptive approach, by questioning how individuals actually behave, how this behaviour is reflected in cognitive processes and how it is effected by neural circuits in the brain.

While studying moral decision-making, many questions necessarily arise concerning why, when and how individuals choose to act in accordance to their moral values – and why, when and how they do not. A different, but related, question concerns how we perceive – and are influenced – by others' behaviour. Human beings are continuously embedded in a socio-cultural environment that feeds them information regarding the appropriateness of and response to their actions. Therefore, it is not difficult to recall instances where others' behaviour influences and spills over ours. For instance: a colleague stole the upcoming high-school mid-term exam from the teacher's desk and shared it with the class. Everyone studied it thoroughly in advance to guarantee a good result. Do you also have a look, or you refrain from doing so? In this scenario, others' dishonest behaviour (i.e., cheating on a test by knowing the answers beforehand) might impact one's own decision to be honest.

Generally, honesty is defined as a qualifying disposition to act according to one's community moral rules and abide by its social norms (Heintz, Karabegovic, Molnar, & Heintz, 2016). In a narrower sense, honesty can be defined as absence of lying and, like other prosocial behaviours, it generally involves the trade-off between self-interest and other (e.g., moral, social) concerns.

The behavioural study presented here is part of a larger computational neuroscience project that intends to integrate functional magnetic resonance imaging (fMRI) and transcranial direct current stimulation (tDCS)¹ data. This project aims to answer whether

¹ tDCS stimulation can influence resting membrane potentials and spontaneous neural activity, sharing features with long-term potentiation of synapses. Depending on current flow direction relative to neuronal orientation, it

observing others' moral behaviour results in changes in behaviour and moral preferences, and the degree of sophistication of contagion. Furthermore, it intends to examine whether these effects are reflected in modulation of brain areas typically involved in moral decision-making and conformity. The focus of the project lies on two processes: learning about group behaviour and contagion from group behaviour. Here, we define learning as the ability to extract what the average behaviour of a group is, which would enable individuals to predict their behaviour. On the other hand, we define contagion as the resulting diffusion of knowledge and the adoption of others' behaviour, or the shift in individual behaviour towards that of others. In essence, it is equated with social learning in its classical sense (Bandura, 1977). Fundamentally, we are treating learning as "learning about", while we treat contagion as "learning from". This distinction is based on the notion that you can learn what others are doing and not consider it behaviourally relevant to your own decisions.

The present report focuses on the second specific goal of the project, behavioural contagion, and presents the following structure: in the first section, we will introduce the topic and discuss moral decision-making, honesty and conformity. We will then proceed to elaborate on the research question and hypotheses. In the second section, we will detail the methodological approach taken to answer the research question. In the third section, we will present our results. The final section will discuss the results in light of the reviewed evidence.

Moral decision-making refers to the choices made in the face of ethical questions or dilemmas, ultimately based on circumstances, personality, preferences, religious beliefs, logical reasoning, amongst other factors. Early psychological studies have investigated how moral behaviour develops over time in infants (Kohlberg, 1981, 1984; Piaget, 1948) and the sort of psychological structures required to incur in moral judgement. One way of studying moral development is the use of stories. Tracing back to Piagetian studies (Piaget, 1948), the latter provide a way to probe into how an individual would behave in a specific situation or how they would perceive the moral character of an action (e.g., Leloup, Dongo Miletich, Andriet, Vandermeeren, & Samson, 2016; Saxe & Powell, 2006; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010; Young, Cushman, Hauser, & Saxe, 2007). A related stream of enquiry relies on the use of moral dilemmas (e.g., trolley problem; Foot, 2002). In these, a subject is confronted with a paradox where two possible solutions (in the form of

can hyper- or depolarize cortical neurons, modulating spontaneous firing frequency. Anodal tDCS has been shown to enhance cortical facilitation, while cathodal tDCS is associated with intracortical inhibition.

moral imperatives) oppose, creating an ambiguous situation in which one prevails over the other.

In more indirect ways, morality can be studied by probing into justice, fairness, equality, equity, altruism, honesty and others, that is, fundamental moral values (Decety & Yoder, 2017; Vermunt, 2016; Wheatley & Decety, 2015). The use of social games to study moral or social behaviour can also be traced back to Piagetian studies and has been picked up by behavioural economists (e.g., prisoner's dilemma; Luce & Raiffa, 1957). Scales and questionnaires have also been used in clinical settings to assess individual behaviour (Kurtin & Pimm, 1983; Rush, First, & Blacker, 2008). Conversely, in laboratory settings, many different paradigms have emerged that attempt to imbue decisions with real-life consequences, instead of relying on hypothetical dilemmas. For instance, the use of electrical shocks to examine the trade-off between profit for self and pain for others (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017; Lockwood, Klein-flügge, Abdurahman, & Crockett, 2019; Volz, Welborn, Gobel, Gazzaniga, & Grafton, 2017) and the use of monetary donations to good- and bad-valenced institutions (Obeso, Moisa, Ruff, & Dreher, 2018; Qu, Hu, Tang, Derrington, & Dreher, 2019; Qu, Météreau, Butera, Villeval, & Dreher, 2019).

The emerging field of social neuroscience has contributed to this stream of research and has unveiled that moral cognition arises from a diverse and decentralised neural network, engaging varied brain structures in a context-dependent manner. Worth mentioning is the involvement of the prefrontal cortex, including areas such as the ventromedial prefrontal cortex (vmPFC), that has been repeatedly implicated in social cognition and has a critical role in encoding the emotional value of sensory stimuli; and of the dorsolateral prefrontal cortex (dlPFC), implicated in cognitive control and deliberate thinking. The temporoparietal junction (TPJ) has also been implicated and is thought to play a role in understanding abstract conceptual representations, perspective taking, belief attribution and intentionality. Furthermore, the anterior cingulate cortex (ACC) has been implicated in emotional responsivity and conflict monitoring, in particular in real decisions; while the posterior cingulate cortex (PCC) is linked to reasoning in hypothetical decisions. Lastly, the amygdala is especially associated with the processing of the harmful consequences of moral judgements and with the processing of moral emotions (see FeldmanHall & Mobbs, 2015; Fumagalli & Priori, 2012 for a comprehensive review of the “moral brain network”), and has been implicated in the escalation of dishonesty over time, mirrored by the amygdala's diminished sensitivity (Garrett, Lazzaro, Ariely, & Sharot, 2016).

As mentioned previously, the focus of the present research lies on honesty in its narrower connotation (i.e., absence of lying). We consider honesty as part of the morality framework in the sense that it confronts a specific moral imperative (telling the truth) with lying for some benefit. Recently, researchers in the disciplines of psychology and economics have started to devote their attention to the topic and a plethora of paradigms have been developed (for a review, see Jacobsen, Fosgaard, & Pascual-Ezama, 2018; Rosenbaum, Billinger, & Stieglitz, 2014).

One of these paradigms is the die-roll task (Fischbacher & Föllmi-Heusi, 2013; Fischbacher & Heusi, 2008)². In this task, subjects report information on a random outcome, where each outcome is associated with a specific payoff. As such, they can increase earnings by reporting higher-payoff numbers³. In an anonymous situation, where information is private, participants are free to misreport the outcome with impunity. Dishonesty is assessed on the aggregate level by comparing the theoretical distribution of outcomes with the actual reported outcomes. For instance, in the case where individuals have to report the outcome of a fair 6-faced die roll, the theoretical average performance is 3.5. However, in a computerised version of this task, it is possible to access individual decisions, making the task observable, at the risk of influencing the participants' behaviour (Gneezy, Kajackaite, & Sobel, 2018). This task typically functions as a positive-sum game, where overreporting does not result in clear higher expenses for anyone involved, apart from the experimenter.

A variation of this game was put forth by Shalvi, Dana, Handgraaf, and De Dreu, (2011), in which participants roll the die more than once and are instructed to report the first roll. In this paradigm, reporting a number that was seen but was not the first one seen is considered a “justifiable lie”, as opposed to a number that was not seen at all. The authors found that people rely on self-justifications to enable themselves to lie and that there is a desire to appear honest in one's own eyes, beyond caring about appearing honest to others (Lelieveld, Shalvi, & Crone, 2016). People who think highly of themselves regarding their moral standards might employ mechanisms that allow them to engage in a limited amount of dishonesty while maintaining a positive self-concept, a theory referred to as the self-concept maintenance theory (Mazar, Amir, & Ariely, 2008). Further, there might be individual differences in how much being honest is relevant to one's self-concept, and stronger honesty-related values have been shown to be related to truth-telling (Dogan et al., 2016).

² Also referred to as die-under-the-cup task or die-in-a-cup task.

³ Note that the payoff and outcome dimension do not necessarily have to be correlated.

A recent meta-analysis (Gerlach, Teodorescu, & Hertwig, 2019) that integrated data from one-shot, fully anonymous and incentivised experiments concluded that, in the die-roll task, the average report was 4.25 (as opposed to the theoretical 3.5). Furthermore, by comparing amount of lying in the die-roll task with other tasks, the authors concluded that people might feel less comfortable overreporting skill-based than chance-based results. Gneezy et al. (2018), using an observable die-roll task, reported that a high fraction (74.85%) of participants who chose to lie also chose to report the highest payoff, thus lying to the maximal extent. They also observed that the lower the actual payoff, the higher the fraction of dishonest reports. Lastly, they reported that, in a non-observable condition, there was a lower number of liars and more partial lying. While in Mazar et al. (2008) partial lying was attributed to self-image concerns, these results suggest that partial lying may mainly rely on social identity concerns, as it is reduced in the non-observable condition⁴. Finally, another recent meta-analysis (Abeler, Nosenzo, & Raymond, 2019) on the die-roll task and others concluded that people tend to refrain from lying to the maximal extent, as even non-maximal outcomes are reported more often than their true likelihood⁵. These authors corroborated Gneezy et al. (2018) finding that the probability of reporting a given outcome weakly increases with its associated payoff. After running their own experiment, the authors concluded that observability significantly decreased reports, as observed in Gneezy et al. (2018). To account for their results, the authors proposed a calibrated utility function in which the main motives for truth-telling were a preference for being seen as honest and a preference for being honest.

Individual differences in honesty behaviour might be related to differences in the activation of the neuronal control network (i.e., ACC, dlPFC and ventrolateral prefrontal cortex, vlPFC) (Greene & Paxton, 2009), in such a way that the involvement of this network may reflect the process of deciding whether or not to lie, independently of the choice made. Furthermore, dlPFC stimulation (by means of anodal tDCS) has been associated with low levels of lying, but only affected individuals who experienced moral conflict (Maréchal, Cohn, Ugazio, & Ruff, 2017), by affecting the trade-off between honesty and self-interest motives in people who were conflicted between these two motives. Further support for the role of dlPFC in promoting honesty comes from a lesion study (Zhu et al., 2014) that

⁴ This follows from the logic that, if people engaged in partial lying to avoid being seen as dishonest by themselves, there should be no difference between the observable and non-observable condition.

⁵ Which would not be the case if, when people lied, they always reported the maximal outcome.

suggested that dlPFC engagement when behaving dishonestly reflects active, but unsuccessful, engagement in control processes.

Social norms can be defined as implicit or explicit rules for acceptable behaviours, values and beliefs, which are partly sustained by other's approval and disapproval (Elster, 1989; Fehr & Fischbacher, 2004) and have been implicated in moral behaviour and dishonesty (see Cialdini & Goldstein, 2004; Park, Goñame, O'Connor, & Dreher, 2017). They are related to one central concept of social psychology, conformity, that is, the alignment between group and one's own behaviour, attitudes or beliefs (Aronson, Wilson, Akert, & Sommers, 2016). Research on conformity has a long-standing tradition in psychology (Asch, 1951, 1956; Bandura, Ross, & Ross, 1961; Sherif, 1936), and related fields have integrated in their theories and models the fact that people tend to conform to social norms. While social neuroscience has mostly focused on studying behavioural contagion in amoral domains, predominantly using ratings of facial attractiveness (e.g., Klucharev, Hytönen, Rijpkema, Smidts, & Fernández, 2009a; Klucharev, Munneke, Smidts, & Fernández, 2011) and Asch-type paradigms⁶ (Berns et al., 2005), behavioural economics has mostly relied on studying norms of fairness and cooperation (e.g., Fehr & Fischbacher, 2003, 2004).

In general, information regarding others' behaviour is a powerful source of social influence, highly effective in changing behaviours (Nolan, Schultz, Cialdini, Goldstein, & Griskevicius, 2008) and preferences (e.g., risk preferences, Chung, Christopoulos, King-Casas, Ball, & Chiu, 2015; Suzuki, Jensen, Bossaerts, & O'Doherty, 2016; temporal discounting, Apps & Ramnani, 2017). This change in behaviour may snow-ball through iterated observation (Krupka & Weber, 2009). Still, people seem more likely to conform to peers' choices if these are equitable (Wei, Zhao, & Zheng, 2013). Various distinctions have been proposed between types of norms and how they affect behaviour (e.g., Cialdini, Reno, & Kallgren, 1990; Krupka & Weber, 2009). One of these was put forth by Toelch and Dolan (2015) and differentiates between informational and normative influences – the former being linked to the fact that people use social cues to acquire information regarding currently adaptive behaviours, and the latter relating to how individuals demonstrate that they belong to a social group. To make adaptive decisions, the individual would update its confidence in one's own behaviour based on others' decisions and the size of the social group. It seems that

⁶ By Asch-type paradigms we are referring to the methodology employed by Solomon Asch in early experiments in social psychology. People participated in a simple visual task (judging the length of a line) along with confederates, that provided an obvious wrong answer (Asch, 1951). Among other things, Asch measured the number of times people conformed to the majority view, even if it was obviously wrong.

people assign more credibility to judgements or decisions made by a larger group (Insko, Smith, Alicke, Wade, & Taylor, 1985; and in a moral domain, Park et al., 2017), but that the group size effect is more evident when the true value of the object is objective (which is not the case in a subjective and ambiguous moral situation) and when the individual is being observed. This suggests that the group size effect is mediated by the dual concern of being right and being liked.

Zooming into the topic of conformity in the honesty domain, dishonesty appears to be more contagious than honesty, at least in an US-American sample, as there might be cultural differences in how much individuals are affected by others' honesty (Innes & Mitra, 2012)⁷. It is also subject to in-group effects (Gino, Ayal, & Ariely, 2009), as has been found for amoral domains (Izuma & Adolphs, 2013), which points to the role of peer influence in immoral behaviour. In Gino, Gu, and Zhong (2009), observing an in-group peer engaging in unethical behaviour increased participants' likelihood of acting unethically themselves, whereas observing an outgroup peer engaging in unethical behaviour reduced participants' likelihood of acting unethically. Compensatory behaviours for in-group transgressions are also more likely than for out-group transgressions, but only in the presence of outgroup members, a phenomenon that might be associated with guilt. That guilt aversion provides a foundation for explaining honesty has been previously suggested (Battigalli, Charness, & Dufwenberg, 2013; Gary, Charness, & Dufwenberg, 2006). Support has been found for the idea that when the saliency of dishonesty increases cheating decreases, by drawing attention to the moral norm (Gino et al., 2009). Conversely, others' immoral behaviour may increase awareness to the fact that behaving immorally is an option. For instance, Fosgaard, Hansen, and Piovesan (2013), using a cheating paradigm, showed that increasing awareness to the possibility of cheating increased the probability of cheating in women, while men seemed to be already aware of that option. Men showed to be more sensitive to peer influence, cheating more when told that peers did so.

How the brain effects conformity behaviour is a topic of extensive research, the careful analysis of which is beyond the scope of this work. Shortly, if low-process, perceptual tasks are used (e.g., rotation of three-dimensional objects in Asch-type experiments), it seems that conformity is effected through a change in perception, directly affecting low-level

⁷ The authors ran a cross-cultural study, with samples from Arizona (USA), California (USA) and Calcuta (India). They found a strong peer propensity for dishonesty, promoting untruthful behaviour. In both countries, they found evidence of contagion, but in the India experiment this contagion went both ways, as individuals were also more frequently honest if exposed to a norm of honesty.

processing areas (Berns, Capra, Moore, & Noussair, 2010; Berns et al., 2005). In subjective valuation tasks (e.g., ratings of music), in turn, the opinions of others might effect immediate changes in value, mediating value signals in the reinforcement learning (RL) circuitry (Campbell-Meiklejohn, Bach, Roepstorff, Dolan, & Frith, 2010), and reward-based learning (Behrens, Hunt, Woolrich, & Rushworth, 2008). In general, it might be the case that the neural mechanisms involved in social conformity are similar to those involved in behavioural adjustment (Wei et al., 2013), as others' behaviour serves as a sort of feedback. Lastly, the dlPFC has been implicated in detecting norm violations (Zinchenko & Arsalidou, 2018), while the right LPFC has been implicated in conformity behaviour (Ruff, Ugazio, & Fehr, 2013; Ruff, 2018).

So far, we have seen that people lie, but not always to the full extent, nor in every circumstance (Abeler et al., 2019; Gerlach et al., 2019; Gneezy et al., 2018), and that they refrain from lying mainly due to the dual concern of being seen as honest by oneself and by others (Abeler et al., 2019; Lelieveld et al., 2016; Park et al., 2017). As such, when there is room for justifying dishonest behaviour, even if the justification appeases only oneself, there is the tendency to lie more (Mazar et al., 2008; Shalvi et al., 2011). When decisions are being observed by the experimenter, or others, people engage less in partial lying, that is, when deciding to lie, they lie to the full extent more often (Abeler et al., 2019; Gneezy et al., 2018). Finally, we have also seen that others' behaviour is a powerful source of information and that people tend to align their behaviour with that of others' (Cialdini & Goldstein, 2004), even in the moral domain (Innes & Mitra, 2012).

The complex topic of conformity and behavioural contagion has already received attention from neuroeconomics, computational neuroscience and behavioural modelling, namely through the use of sophisticated tools such as Bayesian modelling, to answer how it is enforced. For instance, Suzuki et al. (2016) studied contagion of risk preferences, where individuals had to learn and subsequently predict others' choices, besides choosing for themselves. The authors concluded that the shift in participants' behaviour was better captured by a change in risk preferences rather than a change in the subjective probability judgement, a bias for or against the utility of gambling or a bias for or against the choice probability of the gambling option. Using a Bayesian learner model to account for how participants learned about others' behaviour, the authors suggested that individuals use their own preferences as a starting point for learning and then adjust predictions based on others' behaviour, thus utilising one's own preferences as a prior belief. Conversely, Chung et al.

(2015) reported that risk preferences did not change in the face of information about others, but that this served to guide choice.

A related, but behavioural, computational modelling study on how individuals learn about others' prudence, impatience and laziness (i.e., delay, effort and risk preferences) also provided evidence that the influence of others is not simply due to imitation of overt behaviour, as the authors found evidence for changes in cost-susceptibility (Devaine & Daunizeau, 2017). They used a meta-Bayesian approach to construct a model of how people learn from others' choices. In this framework, it is assumed that there is an objective optimal policy for the decision in a specific environment, and that the individual has access to other agents' private knowledge, through noisy reinforcement signals, and updates his policy accordingly. This premise differs from ours, as it is less sensible to assume there is an objectively better moral policy.

In the present project, we intend to study contagion in the moral domain and how individuals learn and incorporate information about others' moral behaviour. Modelling human behaviour can generate novel and testable predictions about the dynamics of decision-making, allowing to observe how it unfolds over time, how past decisions can influence future ones (Crockett, 2016), but also how the decisions of others can influence one's own decisions (e.g., Fehr & Fischbacher, 2003; Khalvati, Mirbagheri, Park, Dreher, & Rao, 2019; Khalvati, Park, et al., 2019; Park, Sestito, Boorman, & Dreher, 2019).

Given the evidence reviewed above regarding honesty and the fact that there is high heterogeneity in how much people lie, both within and among individuals (Gibson, Tanner, & Wagner, 2013), we hypothesise that:

Hypothesis 1. People are partially dishonest, that is, not in every possible circumstance.

Furthermore, we expect people to be able to i) learn what others are doing and predict their behaviour and, in turn, ii) be influenced by their behaviour (Innes & Mitra, 2012; Suzuki et al., 2016), such that:

Hypothesis 2. People learn about others' (dis)honest behaviour, being capable of predicting it.

Hypothesis 3. People are influenced by others' (dis)honest behaviour, that is, there is behavioural contagion in the honesty domain.

Finally, the present study intends to expound how contagion acts, namely whether it enforces a RL mechanism, that updates in the face of social information. As such, we want to test the following:

Hypothesis 4. Contagion is better explained by a dynamic RL process.

We also want to assess whether contagion modifies one's own moral preferences, as found in Suzuki et al. (2016) in the domain of risk preferences, or if it is best explained by a direct conformity bias, as found in Chung et al. (2015). Here, we operationalise moral preferences utilising two parameters: a self-interest and a moral cost parameter; and preference for conformity using an additional parameter. We constructed and tested a set of models, in an attempt to find evidence in support or against one of the following:

Hypothesis 5a. Contagion acts by influencing one's own moral preferences.

Hypothesis 5b. Contagion acts by generating a conformity bias for or against honesty.

As a first step in this project, we utilized the die-roll task in an online computerised experiment. Contrarily to initial planning, which situated participants in a controlled laboratory environment, the experiment was ran online due to COVID-19 pandemic restrictions, as the data was collected during and shortly after confinement measures in France⁸. The use of this task is mainly due to placing no obvious externalities, not depending on ability and allowing for variations in task parameters. Participants saw a die-roll result and, afterwards, were asked to select from two possible alternatives – the true one and a false one – which die they saw. Each die was associated with a specific payoff, and lying was always associated with higher payoff. In two of the three blocks, participants were confronted with the decisions of other individuals, besides deciding for themselves, and had to predict others' behaviour.

⁸ France was on lockdown between the 17th of March and 11th of May, 2020. We found no difference in amount of lying between subjects who were on lockdown and between subjects who were not, as indicated by a Mann-Whitney test ($z = 0.286$, $p = 0.775$) nor any significant ($p < 0.05$) effects when regressing the number of new deaths, total deaths, new cases and total cases on probability of lying (random-effects logistic regression, see Methods for details and Appendix A, Table A1, for results).

2. Methods

2.1. Participants

To test our hypotheses, we recruited 82⁹ participants, 81 of French nationality and one of Algerian nationality (51 female; $M = 25.59$, $SD = 5.05$ years old, in the range 19-36 years old), to participate in an online study, supported by the platform Testable (www.testable.org). Data was collected between the 7th of April and 29th of May, 2020. Participants were recruited through Facebook groups of local universities and activities, as well as Risc mailing list¹⁰. Ethical approval was obtained by local authorities (2018-A01268-47 CPP Sud Est II). We queried 51 participants, out of 82¹¹, on whether they were students and their field of study. From this sub-sample, 20 were students and five were or had been psychology students.

2.2. Trials

The experiment was divided into three blocks and employed the die-roll task. The choice of using the die-roll task relied on three main factors: i) it being a positive sum game, placing no externalities, ii) that does not depend on ability, which could confound the results and, as seen, probably decrease amount of lying, iii) while allowing for variations in task parameters and, thus, being more suited to capture subtle dynamics. Given that the present study is inserted in a larger project that aims to use neuroscientific methods, it is required that the task is observable and that individuals' decisions are traceable to other neuroscientific data. The structure of the experiment was inspired by that of Suzuki et al. (2016), where participants were presented with two types of trials: Self and Predict.

In Self trials, participants were shown the outcome of a traditional 6-face dice roll and were instructed to report its outcome (see Figure 1a for timing of the trial). When reporting, participants were given two choices, which appeared simultaneously on screen: the true draw, that they had just seen, and a wrong one. Participants made their decision by clicking on the image of the outcome they wanted to report. Each outcome was associated with a specific payoff, constant across the whole task, that was exhibited along the image of the draw (i.e.,

⁹ For a non-parametric two-tailed Wilcoxon Signed-Ranks test, with $\alpha = 0.05$ and power 95%, assuming no parent distribution, a sample size of 25 is required for a 0.7 effect size, 63 for a 0.5 effect size and 128 for a 0.3 effect size (G*Power Version 3.1.9.6; Erdfelder, Faul, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007). Effect size refers to a measure of magnitude of the observed effect, here we use three benchmarks: 0.3, 0.5 and 0.7.

¹⁰ Relais d'information sur les sciences de la cognition (<https://www.risc.cnrs.fr>).

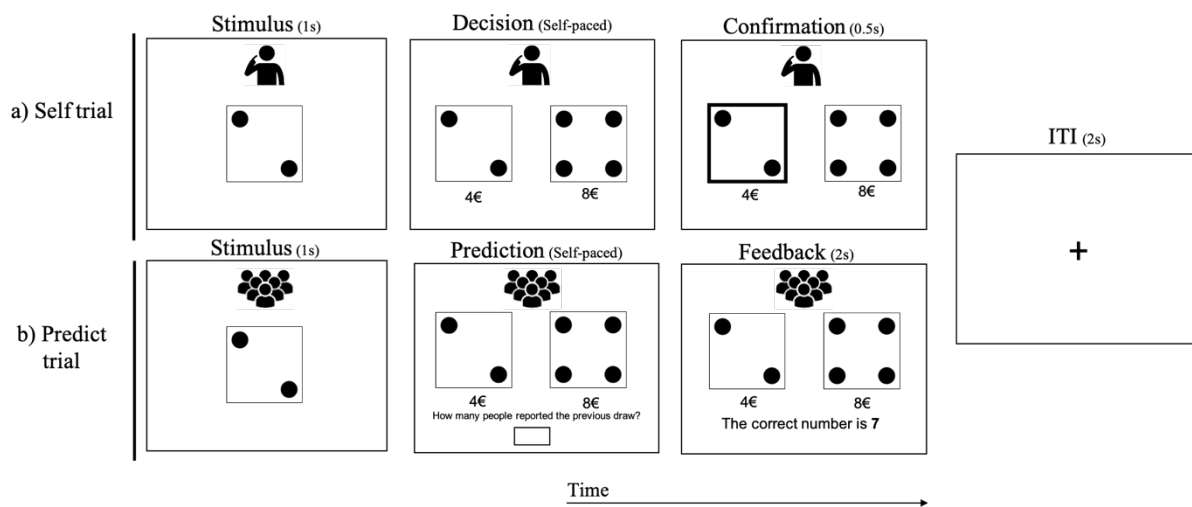
¹¹ The demographic questions were only included in the task after the first participants had already participated in the experiment.

participants saw on screen both the outcome and the payoff associated with it). Participants always had to choose between the draw they saw and a draw associated with a higher payoff. This is, participants faced a dilemma between obtaining a higher reward through lying. The side where the true draw was presented was randomised across trials.

In Predict trials, participants were also shown the outcome of a die roll and saw the reporting screen, with two possible reporting options (see Figure 1b for timing of the trial), as in a Self trial. Instead of making a decision for self, they were instructed to predict how many people of a group of 10 chose to report the draw they had just seen, that is, how many participants reported truthfully. After inputting a number, using their keyboard, participants received feedback on their prediction, as the correct number of individuals who reported truthfully was displayed. The individuals whose actions they had to predict belonged to one of two groups of 10 people (different in block 2 and 3).

Figure 1

Trial structure



Note. a) Self condition. Participants saw the outcome of a 6-faced die roll for 1 second and then were shown a screen with two dices: the draw they previously saw and another, more profitable, draw. They had no time limit to decide which dice to report and, after reporting, they saw, for half a second, a screen confirming their choice. b) Predict condition.

Participants saw the outcome of a 6-faced die roll for 1 second and were shown a screen with two dices: the draw they previously saw and another, more profitable draw. At their own pace, they predicted the number of individuals from a group of 10 that reported the previously seen draw. They then saw the correct number for 2 seconds.

Table 1*Task parameters*

Relative gain from lying	Dice combination				
2€	6:1	1:2	2:3	3:4	4:5
4€	6:2	1:3	2:4	3:5	1:3
6€	6:3	1:4	2:5	6:3	1:4
8€	6:4	1:5	6:4	1:5	6:4
10€	6:5	6:5	6:5	6:5	6:5

Note. Number of the left refers to true draw and on the right to the false draw. In bold are repetitions of task parameters, to accomplish 5 trials per relative gain from lying. For instance, the first cell “6:1” indicates that the participant first sees a 6 (0€) and has the opportunity to report seeing a 1 (2€), which amounts to an additional payoff of 2€ for lying.

Here, each die outcome paid double its value (e.g., the outcome 2 paid 4€), except the outcome six, that paid 0€. These values were chosen as means to ensure the saliency of each choice, while allowing for a no-payoff outcome. The task parameters (i.e., outcomes and associated payoffs) were all possible combinations of two rolls, with additional 15 repeated combinations to achieve a total of 25 trials, resulting in five observations per relative gain from lying (i.e., difference in payoff between reporting dishonestly and honestly; see Table 1). Note that, in this paradigm, being dishonest is always advantageous.

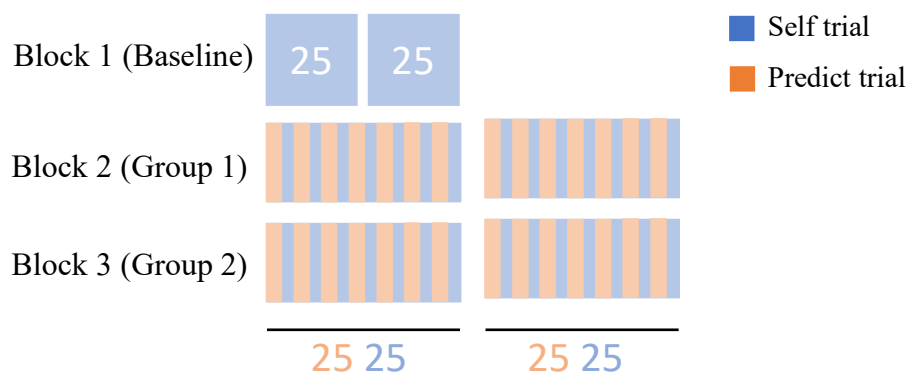
2.3. Task

As mentioned previously, the experiment was divided into three blocks. In block 1, participants were presented with 50 trials of the type Self, separated into two parts of 25 trials. In blocks 2 and 3, trial type was alternated – participants first answered a Predict trial, followed by a Self trial, then another Predict trial, and so on. In each block, they saw a total of 50 trials of each type, also separated into two parts of 25 (see Figure 2). A symbol was present on screen indicating whether the current trial was a Self trial (an icon of a person pointing at himself/herself) or a Predict trial (an icon of a group of people; see Figure 1).

Participants were instructed that they were going to predict the decisions of two distinct groups of individuals in block 2 and 3. The Predict trials refer to the prediction of behaviour of members of the first group, during Block 2, and of the second group, during Block 3 (see Group Behaviour Manipulation section). This instruction was reinforced at the beginning of each block.

Figure 2

Experimental design



Note. Experiment divided in three blocks with Self and Predict Trails. In block 1, participants complete 50 Self trials; in block 2, 50 Self and 50 Predict trials; in block 3, 50 Self and 50 Predict trials.

Two different sequences of trials were created so that participants did not decide on particular task parameters after seeing a Predict trial using that combination of parameters in the previous two trials, for each part, to avoid pure availability and memory effects and force the subjects to learn the norm. Half of the participants completed sequence 1 and half of the participants completed sequence 2 (see Table A2 in Appendix A). No differences were identified between probability of lying on sequence 1 and sequence 2¹².

Before the experiment, participants completed four trials of each type and filled in a comprehension questionnaire, composed of four questions (see Appendix B). Average score on the comprehension questionnaire was 81.10% ($SD = 21.94\%$).

To incentivise every decision, participants were paid at the end of the experiment one randomly selected Self trial (maximum of 10€), plus one randomly selected Predict trial (2€

¹² Hypotheses tested using Mann-Whitney U test ($W = 702.5$, $p = .201$).

if the prediction was correct, 0€ if incorrect), plus a participation fee (1€), and were paid through the platform Testable.

2.4. Group behaviour manipulation

Participants were informed that two distinct groups of 10¹³ other participants had already performed the task in a previous study, without receiving information about others' decisions (i.e., only performed block 1); and that we had selected and separated them into two groups based on their behaviour. Further, they were told that the other individuals had shown homogenous behaviour to other group members, that they were similar to themselves in terms of age and nationality and that they were recruited in the same way. See Appendix B for instructions and comprehension questionnaire, in English¹⁴.

These two groups consisted of an “honest” and “dishonest” group, randomised across participants in order of appearance. As participants had to predict group decisions while receiving feedback, we expected them to learn the average group behaviour and, thus, the social norm prevalent in each group. This norm was either an honest norm, where most group members chose to report truthfully, or a dishonest norm, where most group members chose to lie. From here on, we will refer to the block where participants predicted the behaviour of the honest group as the honest block, and to the block where participants predicted the behaviour of the dishonest group as the dishonest block.

The decisions of the group were simulated utilising a model which assumes that the difference in subjective utility between lying and being honest relies on a self-interest α parameter, which weighs the relative gain from lying; and a moral cost δ parameter, which weighs the absolute gain from lying. This model is described in detail in the section Analysis, under the name Basic Model. We ran the simulation on MatLab (R218b; MATLAB, 2018), using the Variational Bayesian Analysis (VBA) toolbox (Daunizeau, Friston, & Kiebel, 2009; Daunizeau, 2020; Daunizeau, Adam, & Rigoux, 2014), using the following parameters: for the honest group, α was set to 0.11, δ to -1.12 and β (temperature in the softmax function) to 1.24; whereas for the dishonest group, α was set to 0.72, δ to -1.10 and β to 1.26. Note that the value of α , the self-interest parameter, is lower for the honest group, which translates into a lower likelihood of reporting the highest paying dice at the cost of lying. See Table A3 in

¹³ We chose to report the decisions of 10 individuals instead of a single individual's because bigger groups tend to generate higher levels of conformity (Bond, 2005; Insko et al., 1985; Park et al., 2017)

¹⁴ Freely translated from the original French.

Appendix A for specific number of honest and dishonest choices associated with task parameters and block in Predict trials.

These parameters were chosen based on an on-site pilot study ($N = 11$) run in March 2020 at the Institut des Sciences Cognitives Marc Jeannerod (Lyon, France). We selected two subjects whose behaviour and average lying fitted the desired for the honest and dishonest group and used their estimated parameters, again using the aforementioned Basic Model¹⁵, to simulate the behaviour of 10 individuals. In this pilot study, different task parameters were used, that resulted in very low probability of lying ($M = 0.130$, $SD = 0.337$), thus we chose different task parameters¹⁶ for the present study and adjusted the results of group behaviour estimates accordingly. Hence, the study resorts to deception, as the honest and dishonest group's decisions are experimentally manipulated, unbeknownst to the participants. It should be noted, as previously explained, that this behavioural study is preliminary to an fMRI study, where the use of deception is common.

2.5. Analyses

Analyses were performed in three steps: i) group comparisons, ii) regression analysis and iii) modelling. These analyses served not only the purpose of answering our specific hypotheses, as stated in the Introduction, but also an exploratory one, to allow for a better understanding of the data, as this experimental design and paradigm are a first attempt at characterising moral contagion in the honesty domain. As the focus of the present report lies on the contagion effect, the analyses conducted on learning to predict group behaviour are complementary.

Group comparisons were performed in RStudio (Version 1.1.453) and regression was performed on Stata (Stata/IC 16.1), using a random-effects logistic regression model (function *xtlogit*) and random-effects linear regression model (function *xtreg*), with error clustered at the subject level. The random-effects logistic regression model uses probability of lying as dependent variable and includes various regressors of interest, controlling for demographics and other variables¹⁷. These regressors of interest were chosen based on our hypotheses and consist of the block the participant is in (baseline, honest and dishonest

¹⁵ This model was the one that provided the best fit in the sample and paradigm mentioned.

¹⁶ Previously, each die paid precisely its value (e.g., a roll of 1 paid 1€) and participants were paid a participation fee of 20€. To increase task gains and saliency, we opted for paying double of die value and including the scenario where being honest pays nothing (i.e., a roll of 6 pays 0€), while paying 1€ of participation fee.

¹⁷ Sex, age, sequence, comprehension questionnaire score and time.

block), a dummy for whether the participant faced the honest or the dishonest block first, and task parameters, that is, payoff for reporting true draw and false draw. The random-effects linear regression model uses absolute prediction accuracy (i.e., absolute difference between participant's prediction and correct number of honest people) as dependent variable and includes the same regressors of interest as the random-effects logistic regression model, in addition to part of the block, time, age and questionnaire score, controlling for sex and sequence.

As for modelling, we followed an explorative approach, testing 20 different models using MatLab (R218b; MATLAB, 2018) and the aforementioned VBA toolbox (Jean Daunizeau, 2020).

2.5.1. Decision model and Basic Model

The probability of participant i lying on trial t was computed using the softmax function:

$$P_i(\text{lie}) = \frac{1}{1 + e^{-\beta \Delta U_{i,t}}} \quad (1)$$

$$\Delta U_{i,t} = U_{i,t}(\text{dishonest}) - U_{i,t}(\text{honest}) = \alpha_i(\pi_{D,t} - \pi_{H,t}) - \delta_i \pi_{D,t} \quad (2)$$

Where $\pi_{D,t}$ and $\pi_{H,t}$ refer to the payoff at trial t when being dishonest and honest, respectively. Here, α is a self-interest (greed) parameter, weighing the relative gain derived from lying, i.e., the difference in payoff between the dishonest and honest option, and is expected to be positive; whereas δ is a moral cost parameter, weighing the absolute gain derived from lying, and is expected to be negative, as indicated by the negative sign in the utility function. This utility function is inspired by the model of Zhu et al. (2014)¹⁸. We refer to the model specified here as the Basic Model.

2.5.2. Variations of the Basic Model

Two different main classes of models were tested: static and dynamic models. Dynamic models are updated in a RL fashion, while static models are not. In dynamic models, parameters of interest (e.g., self-interest parameter, α) update trial-by-trial based on group behaviour in Predict trials, weighed by a learning rate. Another important variation

¹⁸ The authors used computational modelling to arrive at a characterisation of the relative contributions of self-interest, social preferences (of money allocations) and honesty to utility and, ultimately, choice. We chose this utility function because it was computationally tractable and suited our 2-choice paradigm.

tested refers to whether parameters of interest were estimated for the whole task or whether these were estimated per block, from here on referred to as standard and multiblock¹⁹ models, respectively. This resulted in a total of 20 models tested: 3 static-standard, 7 static-multiblock, 5 dynamic-standard and 5 dynamic-multiblock (see Appendix C for complete description of models tested).

Three families of static models were tested: the aforementioned Basic one; a Fixed Moral Cost model where δ is independent of $\pi_{L,t}$ ²⁰; and a Conformity model, where a conformity parameter, γ , is added to the equation, independent of task parameters and group behaviour, on top of the self-interest and moral cost parameter²¹. While the basic and fixed cost model assumes a change in individuals' moral preferences (self-interest and moral cost), the conformity model assumes these are fixed, and in turn introduces a change in the preference to conform, which in turn creates a bias towards or against lying in the face of social information.

Conversely, two families of dynamic RL models were tested: a Moral and a Conformity family. In the Moral family, α , δ or both are updated as a function of the amount of truth-tellers in the previous Predict trial according to a learning rate²². In the multiblock version of these models, the learning rate and the hidden states are estimated separately in the honest and the dishonest block. In the Conformity family, the conformity parameter γ updates according to a learning rate. In the multiblock version of these models, the learning rate and the hidden states are estimated separately in the honest and dishonest block. The Conformity family is divided into two types: a Conformity-Action type, that updates as a function of amount of truth-tellers in the previous Predict trial; and a Conformity-Outcome type, that updates as a function of both amount of truth-tellers and relative gain derived from lying (i.e., difference in payoff between reporting the false draw and the true draw) in the previous Predict trial.

In general, dynamic RL models proved to be inadequate to explain individuals' behaviour, seemingly unfit due to a low amount of contagion and lack of contagion dynamics. This is the main reason why more sophisticated models of learning²³ and contagion (e.g., Bayesian learner models) were not explored in this report.

¹⁹ The temperature parameters, β , is assumed to be fixed throughout. Only parameters of interest are estimated per block.

²⁰ In the multiblock versions of the Basic and the Fixed Cost model, α , δ or both are estimated per block.

²¹ In the multiblock version of this model, γ is estimated per block.

²² These models are referred to as Moral- α , Moral- δ and Moral- α - δ .

²³ Note that, in the present reinforcement learning models, individuals' predictions were not used, as we did not use a prediction error in its classical sense.

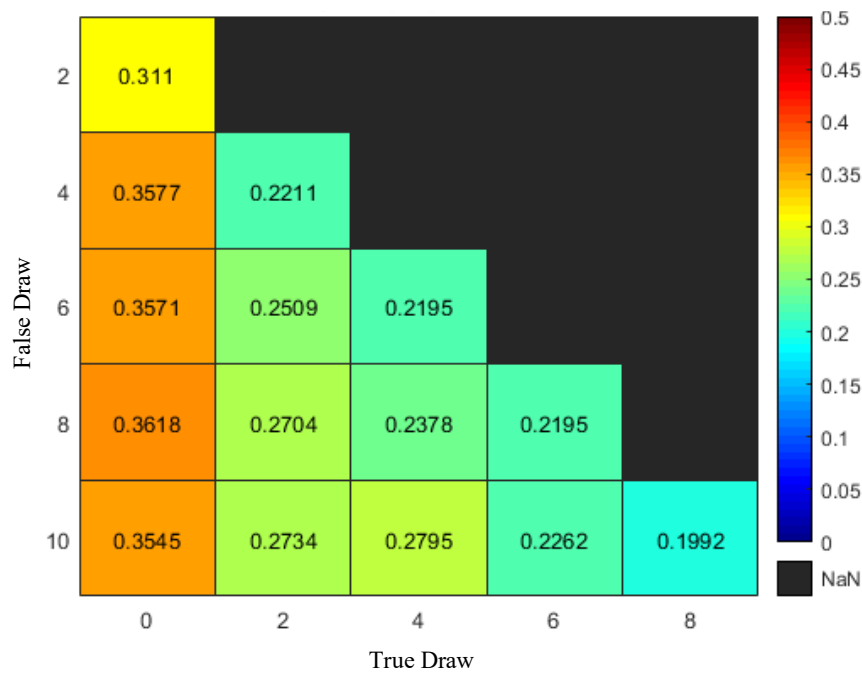
3. Results

3.1. Overall lying

Among the 82 participants, 12 were fully honest ($N = 9$) or fully dishonest ($N = 3$), that is, they told the truth or lied in every Self trial, while the others exhibited heterogenous behaviour. The overall probability of lying was modest but varied greatly (29.75% of lies across all Self trials and subjects, $SD = 45.72\%$). It also varied with task parameters. From visual inspection of Figure 3, in which we see a heatmap of probability of lying as a function of task parameters, it can be seen that individuals seem sensitive to both the value of the true draw and the alternative (false) draw (see Figure A1 in Appendix A for heatmap of each block). These results partially support hypothesis 1, as 85.37% ($N = 70$) of the participants lied at least once or were truthful at least once²⁴.

Figure 3

Probability of lying as a function of task parameters (true draw and false draw)



Note. Task parameters refer to the payoff associated with the true and false draw.

²⁴ No significant differences were found in overall lying between men and women, as indicated by a Mann-Whitney test ($U = 923.5$, $p = .204$).

We thus found that:

Result 1. Most people are partially dishonest and do not lie in every possible circumstance.

Further, we saw that lying resulted in slower reaction times, as indicated by a Wilcoxon Signed-Ranks test on individual reaction times²⁵ ($T = 2095$, $p < .001$, $r = 0.40$ ²⁶).

3.2. Learning

We found evidence that participants learned to predict the behaviour of each group, in such a way that we found a difference between predictions in the honest block ($Mdn = 8.9$) and dishonest block ($Mdn = 2.5$), as indicated by a significant Wilcoxon Signed-Ranks test on individual predictions²⁷ ($T = 2926$, $p < .001$, $r = 0.591$ ²⁸). This result lends support to hypothesis 2, by signalling that participants are capable of differentiating between the behaviour of the two groups, as well as predict it.

A sequence of Wilcoxon Signed-Ranks tests on prediction accuracy (i.e., difference between participant's prediction and true number of people who reported truthfully), corrected for multiple comparisons with the Holm-Bonferroni method (Holm, 1979), suggests that, in the honest block, the accuracy in the first part of the block was higher than in the second ($T = 2929.5$, $p < .001$), the latter not differing significantly from zero ($T = 1484.5$, $p = .913$), while the former did ($T = 2906$, $p < .001$). Conversely, in the dishonest block, the accuracy in the first part was smaller than in the second ($T = 806.5$, $p < .001$), the latter not differing significantly from zero ($T = 1272$, $p = .364$), while the former did ($T = 595.5$, $p < .001$). Furthermore, during the first part, the accuracy in the honest block was significantly

²⁵ Reaction times were calculated on an individual level for trials where participants were honest and dishonest and then subject to statistical analysis. We chose to report non-parametric tests because reaction times deviated significantly from normality, as indicated by Shapiro Wilk normality tests (dishonest: $W = 0.7071$, $p < .001$; honest: $W = 0.5752$, $p\text{-value} < .001$).

²⁶ $z = 4.989$ and effect size of $4.989/\sqrt{164} = 0.40$, indicating a medium effect. z -score calculated using function `wilcoxsign_test` from package `coin` in R. Effect size calculated as indicated in Field (2013). Cohen's criteria of 0.3 and 0.5 for a medium and large effect are used.

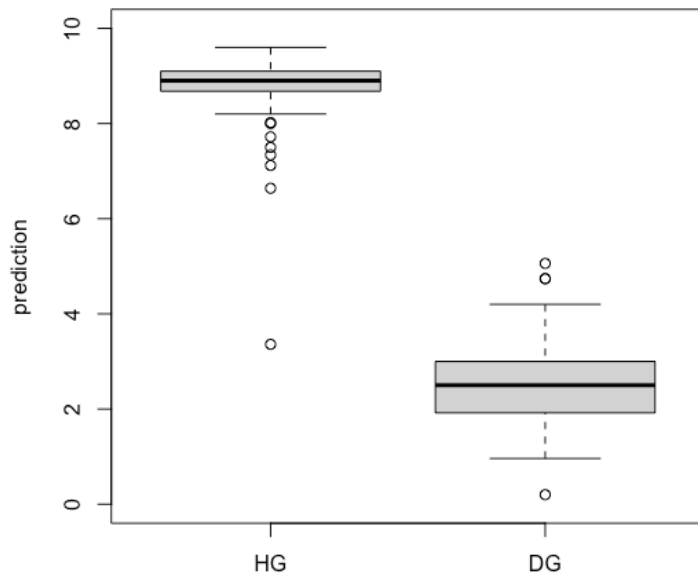
²⁷ Predictions were calculated on an individual level for the honest and dishonest block and then were subject to statistical analysis. We chose to report non-parametric tests because predictions of honest group's behaviour deviated significantly from normality, as indicated by Shapiro Wilk normality test ($W = 0.61223$, $p < .001$).

²⁸ $z = 7.5746$ and effect size of $7.5746/\sqrt{164} = 0.591$, indicating a large effect. See note 26 for details on effect size calculation.

greater than that in the dishonest block ($T = 2974.5$, $p < .001$), but there was no evidence for a difference in the second part ($T = 1858.5$, $p = .341$).

Figure 4

Predictions on Predict trials when facing the honest and dishonest groups



Note. Participants were instructed to predict how many individuals from a group of 10 reported the previously seen die roll (i.e., how many were honest). HG = Honest Group block; DG = Dishonest Group block.

This suggests the following: i) in the first part of the honest block, the predictions are underestimated, that is, subjects are expecting lower amount of truth-tellers, resulting in positive accuracy deviations, ii) in the first part of the dishonest block, predictions are overestimated, that is, subjects are expecting higher amount of truth-tellers, resulting in negative accuracy deviations; and iii) in the second part of both blocks, the norm of each group is thoroughly learned.

To explore further differences on how individuals learned to predict group behaviour, we ran a random-effects linear regression model on absolute prediction accuracy (i.e., absolute difference between participant's prediction and correct number of honest people), with various regressors of interest (see Table 2).

Table 2*Results of Random-Effects linear regression analysis on absolute prediction accuracy*

Predictor	Coefficient
Intercept	8.165*** (0.574)
Age	0.036** (0.012)
Questionnaire Score	-0.843** (0.281)
Time	-0.048*** (0.003)
Second Part	0.602*** (0.099)
Dishonest Block First	-3.057*** (0.217)
Honest Block	-3.966*** (0.184)
True Draw	0.149*** (0.011)
False Draw	-0.148*** (0.011)
Dishonest Block First × Honest Block	5.738*** (0.356)

Note. Standard error, clustered at the subject level, is reported in parentheses (***p < 0.01; **p < 0.05; *p < 0.1). $\chi^2(11) = 1326.81$, prob > $\chi^2 = 0.000$, Overall $R^2 = .147$.

As already described, we found a significant decrease in absolute prediction accuracy in the second part of each block, when compared to the first, of 0.602 ($p < .001$). We also found a significant effect of being in the honest block ($p < .001$), which resulted in a decrease of 3.966 in absolute prediction accuracy. This indicates that participants are better at predicting the decisions of the honest group than those of the dishonest group. Furthermore, having seen the dishonest block first decreased absolute prediction accuracy by 3.057 ($p <$

.001), suggesting that learning about the dishonest group first rendered individuals better at predicting other's decisions. Having seen the dishonest block first and being in the honest block, together, result in an additional increase of 5.738 ($p < .001$) in absolute prediction accuracy. This implies that the beneficial effect of being in the honest block in the ability to predict choices fades if the first group whose behaviour was learned is the dishonest one.

We also found significant effects of both the payoff for the true draw ($p < .001$) and false draw ($p < .001$), of similar magnitude, in different directions. Increasing true draw increases absolute prediction accuracy by 0.149, while decreasing false draw decreases absolute prediction accuracy by 0.148²⁹. Note that, in this context, true and false draw refer to the dice seen in the Predict trial; and not the ones that the participant sees when deciding for oneself. This result suggests that higher values of true draw render individuals less able to predict other's decisions, while lower values of false draw make their decisions more predictable. Why this happens is an open question but might be related with the variability in group behaviour itself.

Lastly, we found a significant positive effect of age ($p = .003$), suggesting older individuals made worse predictions; a negative effect of questionnaire score ($p = .003$), suggesting that individuals who scored higher in the initial comprehension questionnaire made better predictions; and a negative effect of time ($p < .001$), suggesting that, as the task progressed, individuals became better at making predictions.

In summary, we found that:

Result 2. People are capable of predicting others' behaviour.

3.3. Contagion

Now that we have established that participants indeed learned what others were doing, we will delve into the main focus of the report, which is whether others' behaviour influenced one's own. First, we found evidence for differences in amount of lying between blocks (Figure 5). A Friedman test³⁰ of differences among repeated measures was conducted

²⁹ It should be pointed out, for better interpretation of results, that in the present design, the values of draws are increased in steps of 2€. Thus, the effect of increasing true draw one level corresponds to an effect of 0.298 and of increasing false draw to an effect of -0.296 in absolute prediction accuracy.

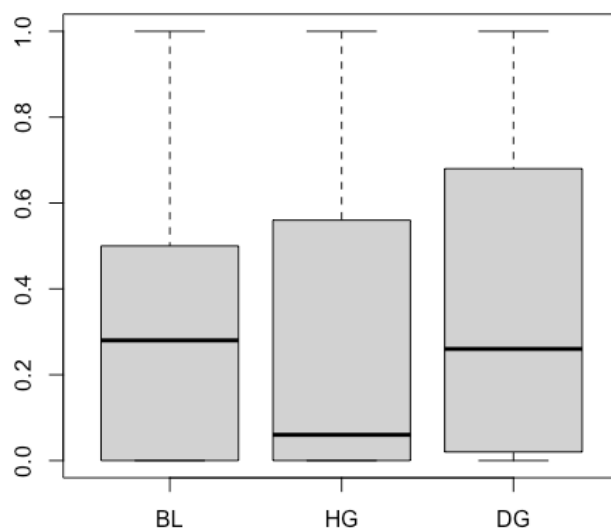
³⁰ Probability of lying was calculated on an individual level for each block and then subjected to statistical analysis. Even though the sample size is fairly big ($N = 82$), we chose to report non-parametric tests for the distribution of participant's lies was far from normal, as indicated by a sequence of Shapiro-Wilks normality tests (baseline: $W = 0.75695$, $p < .001$; honest group: $W = 0.71881$, $p < .001$; dishonest group: $W = 0.76745$, $p < .001$).

on probability of lying and rendered a Chi-square value of $\chi^2(2) = 26.289$ ($p < .001$). A sequence of two-sided Wilcoxon Signed-Ranks, corrected with the Holm-Bonferroni method for multiple comparisons, pointed to a difference between baseline and the dishonest block ($T = 485$, $p = 0.001$, $r = -0.303^{31}$) and between the honest and the dishonest block ($T = 1326$, $p < .001$, $r = -0.359^{32}$), which lends support to hypothesis 1.

We also saw that, for individuals whose behaviour showed some heterogeneity, that is, that were not fully honest nor dishonest, amount of contagion³³ was not correlated with probability of predicting correctly ($r = 0.07$, $p = .595$). This supports the notion that contagion was not primarily triggered by ability to predict choices, thus supporting the dissociation between learning and contagion, as defined in this work.

Figure 5

Probability of lying in each block



Note. BL = Baseline; HG = Honest Group block; DG = Dishonest Group block.

³¹ $z = -3.8778$ and effect size of $-3.8778/\sqrt{164} = -0.303$, indicating a medium effect. See note 26 for details on calculation.

³² $z = -4.5948$ and effect size of $-4.5948/\sqrt{164} = -0.359$, indicating a medium effect. See note 26 for details on calculation.

³³ Amount of contagion was defined as deviations from the previous block: a positive deviation indicating a tendency to conform and a negative one to anti-conform. For instance, if the individual saw the dishonest group first, we expect higher probability of lying when facing the dishonest group as compared to baseline, and a smaller probability of lying when facing the honest group as compared to the dishonest group. It should be noted that for extreme behaviour, i.e., more extreme than the norm to be learned, this logic might not apply. In this case, if the individual is more honest at baseline than the honest norm, then contagion would imply that he becomes more dishonest when facing the honest group. The same applies to the situation where he is more dishonest at baseline than the dishonest norm.

Table 3

Marginal effects associated with the Random-Effects Logistic Regression analysis on probability of lying

Predictor	Dy/dx
Dishonest Block First	.146** (.067)
Honest Block	-.017 (.019)
Dishonest Block	.037* (.074)
True Draw	-.034*** (.007)
False Draw	.011*** (.003)

Note. Standard error, clustered at the subject level, is reported in parentheses (*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$). $\chi^2(11) = 59.36$, $\text{prob} > \chi^2 = 0.000$.

To look further into these results, we ran a random-effects logistic regression with the probability of lying as dependent variable, subject as the group variable and various regressors of interest (see Table 3).

We found an effect of seeing the dishonest block first, which resulted in a robust ($OR = 5.876$; see Table A4 in Appendix A) increase of 0.146 in the probability of lying ($p = .029$). This suggests that there is a priming-like effect on individuals' behaviour. Being in the honest block did not seem to affect amount of lying, when compared to baseline ($p = .368$), whereas we found a trend for an increase of 0.037 on the probability of lying ($OR = 1.574$, Table A4 in Appendix A) when being in the dishonest block ($p = .074$), as compared to baseline.

Both the payoff of the true draw and false draw seem to have an effect in probability of lying: an increase in the payoff of the true draw resulted in a decrease of 0.034 in probability of lying ($p < .001$), whereas an increase in the payoff of the false draw resulted in an increase of 0.011 in probability of lying (by 0.011, $p < .001$)³⁴. Furthermore, these two

³⁴ As mentioned previously, in the present design the payoff from draws are increased in steps of 2€. Thus, the effect of increasing the outcome of the true draw one level corresponds to a decrease of 0.068 in probability of

seem to interact (see Figure 6 and Table A4 in Appendix A), as increases in the payoff of the false draw have a higher effect in amount of lying when true draw is higher ($p = .002$). No other effects emerge as significant³⁵.

In summary, we found:

Result 3. People are slightly influenced by others' dishonesty, resulting in an increase of 14.6% in probability of lying as compared to when no social information is being learned.

In sum, we saw that contagion acted in such a way that people lied more when they were learning about others' dishonesty than without social information, lied more when they were learning about others' dishonesty than when learning about others' honesty and, overall, learning about the dishonest group first was associated with higher probability of lying.

3.4. Model-based analysis

Even though we found a limited contagion effect, we explored the results further via modelling, in an attempt to capture how contagion unfolded. The models estimated were inserted in a group-level random-effect Bayesian model comparison, which treats models as random effects that could differ between participants with an unknown population distribution, instead of imposing the same model for every participant. Given that the purpose of the model-based analysis was to capture behavioural contagion, subjects whose behaviour was homogenous (fully honest, $N = 9$, and dishonest subjects, $N = 3$) were excluded from further analysis.

In the resulting sample ($N = 70$), we found that the Conformity model where parameters of interest are estimated per block (model 10 in Figure 7 and in Table C1 in Appendix C) was the best suited to explain most subjects' behaviour ($N = 48$, 69%, see Figure 7), scoring an exceedance probability of 99.75%. Exceedance probabilities correspond to the posterior probability that a given model is the most frequent in the sample (Devaine &

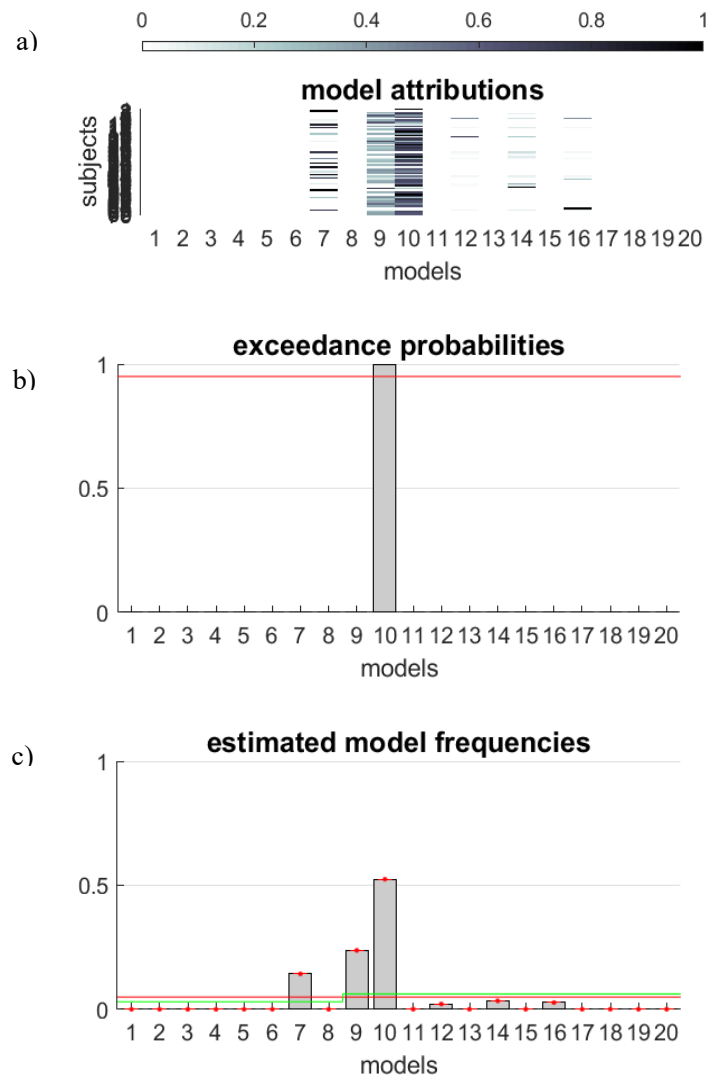
lying; and of increasing the outcome of the false draw to an increase of 0.022 in probability of lying. The same is applied to the interaction.

³⁵ This model was tested for fit against other models using the Bayesian Information Criterion and according to significance of effects, without the non-significant effects of demographic variables. These were included in the final model for completeness. When looking solely at the 51 participants whom we queried whether they studied psychology and whether they were students, the only effects that survive is the effect of seeing the dishonest group first (by 0.178, $p = 0.041$) and the value of the true draw (by -0.025, $p = 0.008$), perhaps due to lack of statistical power. Including the two additional demographic variables did not change these effects. We chose to focus on the complete data set to answer our questions, as we did not have specific hypothesis concerning these demographics.

Daunizeau, 2017). Bayesian omnibus risk (BOR) was virtually zero ($1.27e^{-10}$). This value indicates the posterior probability that model frequencies are equal, and its low value suggests that it is extremely unlikely that these results are driven by chance (Devaine & Daunizeau, 2017). Nonetheless, this model also scored a mean classification accuracy of 57.87%, which is low taking into account that the choice is binary (probability of hitting at random of 50%) and scored a mean R^2 of 0.159. The accuracy of the model being low is not surprising, given the low levels of contagion found.

Figure 7

VBA output



Note. a) Model attributions indicate likelihood of each model for each subject. b) Exceedance probabilities correspond to the posterior probability that a given model is the most frequent

one, indicating that the most likely model is the Conformity multiblock model. c) Estimated model frequencies \pm posterior standard error.

Nonetheless, some individual's behaviours were best explained by other models, namely the Fixed Cost- δ model ($N = 11$; $\sim 16\%$), the Conformity model ($N = 7$; 10%), the Moral- α multiblock model (12) ($N = 2$; $\sim 3\%$), the Moral- δ multiblock model ($N = 1$; $\sim 1\%$) and the Moral- α - δ multiblock model ($N = 1$; ~ 1). The Fixed Cost- δ model assumes that the moral cost δ is independent of task parameters and is estimated per block, while the self-interest α , a positive parameter, that weighs the relative gain from lying, is estimated for the whole task. The standard Conformity model assumes that α , δ and γ are fixed throughout the task, but that γ , the conformity parameter, is null during baseline. The Moral RL models assume that α , δ or both are updated according to amount of truth-tellers in the previous Predict trial; and the fact that they are multiblock indicates that the learning rate is estimated separately for each block (see Appendix C for a detailed description). Given the low prevalence of these models and the high exceedance probability found for the Conformity multiblock model, we focused on analysing the results in light of this model.

In general, we did not find evidence in favour of hypothesis 4:

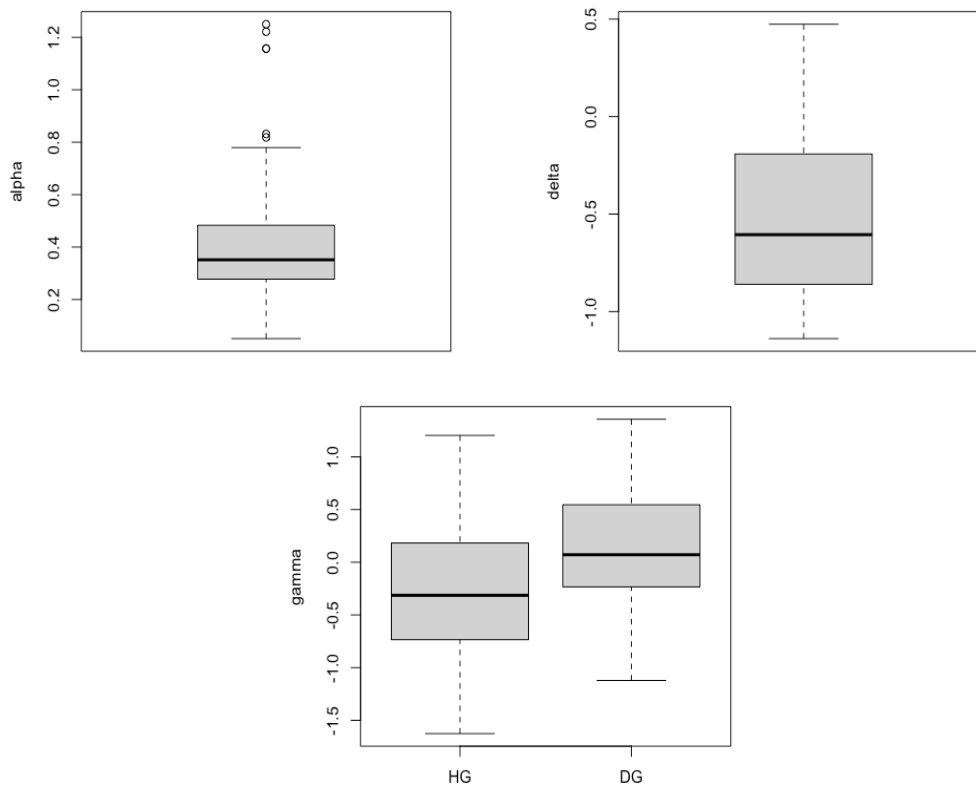
Result 4. A dynamic RL process is not adequate to capture most individuals' behaviours.

In turn, we found evidence in support of hypothesis 5b and found weak evidence in favour of hypothesis 5a, as for most participants behaviour was not explained by changes in moral preferences, as defined by the self-interest α parameter and the moral cost δ parameter. Instead, participants exhibited a preference for conformity, reflected in the γ parameter, independent of task parameters, which introduces a conformity preference towards or against being dishonest in the face of social information.

Result 5. For most participants, contagion acted by introducing a preference bias for or against honesty.

Figure 8

Distribution of parameters in the Conformity multiblock model



Note. α is a self-interest parameter, δ a moral cost parameter, and γ a conformity parameter. Please note the scaling. HG: Honest Group block; DG: Dishonest Group block.

Looking closely at the parameters estimated for the Conformity multiblock model (Figure 8), where only the conformity parameter γ is allowed to vary between blocks, it can be seen that, when participants are in the honest block, the conformity parameter becomes negative, thus decreasing the value of lying compared to being honest; and becomes positive in the dishonest block, thus increasing the subjective value of lying compared to being honest.

Comparing the values of γ in the dishonest and the honest block, we find that the values in the honest block are smaller than the values in the dishonest block, as indicated by a significant Wilcoxon Signed-Ranks test ($T = 405$, $p < .001$, $r = -0.376^{36}$). While the honest block values differ significantly from 0 (Wilcoxon Signed-Ranks test, $T = 661$, $p < .001$),

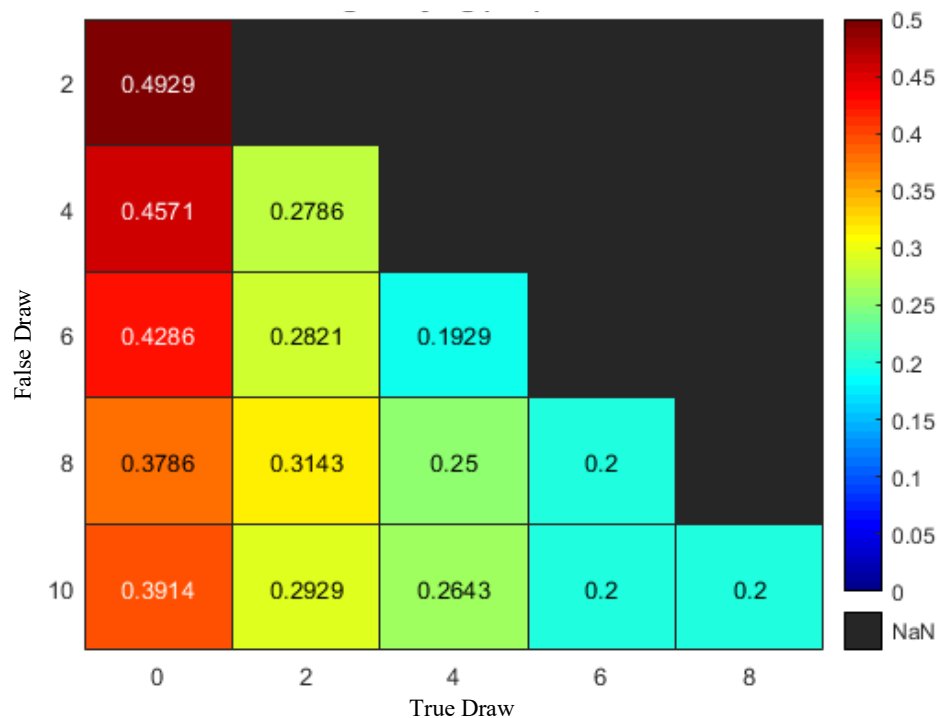
³⁶ $z = -4.8106$ and effect size of $-4.8106/\sqrt{164} = -0.376$, indicating a medium effect. See note 26 for details on calculation.

there is only a trend for the same effect in the dishonest block (Wilcoxon Signed-Ranks test, $T = 1540$, $p = .082$)³⁷.

Using the values of α , δ and γ estimated for each individual and simulating the behaviour of the 70 individuals, we find a slightly higher probability of lying (31.73%) than in the real data, as well as big variations in the amount of lying according to task parameters ($SD = 46.55\%$, Figure 9; and see Figure A2 in Appendix A for heatmap of each block). This result is not far from what we find in our sample: the direction of the gradient and sensitivity to task parameters converge with our empirical results, even if the effect of task parameters is more pronounced in the simulation (see Figure 10). Nonetheless, the elevated probability of lying when the false draw was 2 and the true draw was 0 is absent in our data. This suggests that, even though we found a small contagion effect, as indicated by logistic regression, the present model is able to capture relevant aspects of behaviour, thus being a proxy for the mechanisms behind contagion, but it is not flawless.

Figure 9

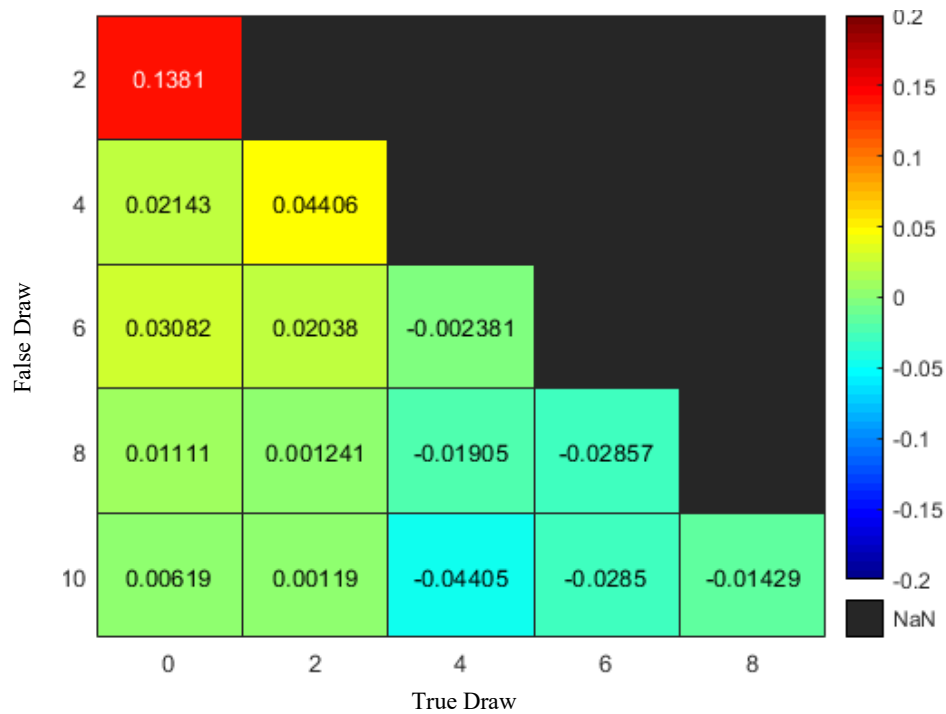
Probability of lying as a function of task parameters, as simulated



³⁷ p-values corrected for multiple comparisons with Holm-Bonferroni method.

Figure 10

Difference in probability of lying between simulation and real data as a function of task parameters



Note. Positive values indicate an overestimation of probability of lying by the simulation, while negative values indicate underestimation.

4. Discussion

The present behavioural experiment aimed at studying contagion in the moral domain, in particular how individuals learn and adopt others' (dis)honest behaviour. We found that most people show heterogeneous behaviour, thus not being honest nor dishonest in every circumstance. This means that most individuals were sensitive to variations in how profitable it was to lie and adapted to it, thus lending support to our first hypothesis, that people are partially dishonest. We also found that the behaviour of the two groups was learned (i.e., was predicted) fairly easily, confirming our second hypothesis, but the behaviour of the honest group was learned more easily than that of the dishonest group. Furthermore, we found that, in general, people are influenced by others' dishonesty, but to a small extent. We also found an order effect, such that being confronted with the dishonest group's decisions first increased overall dishonesty levels. Lastly, through behavioural modelling, we found that the best fitting model that we tested was a static one, without a RL rule. In this model, contagion acts by generating a conformity preference towards or against lying, independent of one's own moral preferences (defined by self-interest and moral cost), that are not best described by a dynamic RL mechanism. This lends support to hypothesis 5b and disempowers hypothesis 4 and 5a.

We found that, overall, individuals were dishonest in 29.75% of the trials, a finding consistent with previous findings in the die-roll task. In Gerlach et al. (2019), the authors report that 30% of reports are over the expected average claim that would result from honest reporting, in fully anonymous (non-observable) settings. Here, we would expect lower levels of lying, given that observability should decrease reported outcome in a free choice situation (Abeler et al., 2019), amount of liars and of partial lies (Kajackaite & Gneezy, 2017), but also that repeated decisions should lower reports (Abeler et al., 2019). However, it should be noted that, even if the participants knew that their decisions would be later analysed by the experimenter, they performed the task alone at home, instead of in a laboratory setting, which might generate a higher sense of anonymity to their behaviour. Our finding that dishonesty was associated with higher reaction times converges with the notion that, when behaving dishonestly, individuals engage in active but unsuccessful control processes (Zhu et al., 2014). This finding is at odds, however, with an intuitive-dishonesty framework, where thinking fast would amplify self-interest in the absence of clear externalities (Kobis, Verschuere, Bereby-Meyer, Rand, & Shalvi, 2019).

Probability of lying varied greatly, reflecting that participants were sensitive to task parameters, but mainly the fact that we observed that some of them ($N = 12$) were fully honest or dishonest, even if the majority showed heterogeneous behaviour. This heterogeneity in behaviour is not surprising and has been accounted for before (Gibson et al., 2013). We found that both the value of the true draw and of the false draw affected rates of lying in such a way that a higher true draw decreased the probability of lying. It has been described before that an increase in the payoff of the true draw decreases rates of lying (Abeler et al., 2019; Gneezy et al., 2018). In the present paradigm, the subject was forced to choose between two possible reports, and we saw that a higher false draw increased rates of lying, that is, participants felt greater temptation to lie when the possible alternative was higher. This is consistent with the finding that, when people decide to lie, a high fraction report the highest payoff (Gneezy et al., 2018). Furthermore, the value of true and false draw interacted, as an increase in the false draw had a higher effect in amount of lying when true draw was higher, suggesting that people are sensitive to the relative payoff gained from lying (i.e., the difference in profit between being dishonest or honest). This might be due to the fact that bigger differences reflect a deviation from reality and willingness to lie for a gain decreases with how much reality has to be distorted (Hilbig & Hessler, 2013).

Lastly, we found that individuals exhibited some behavioural contagion in the moral domain in such a way that we found a tendency to lie more when they were in the dishonest block, a finding consistent in both group comparisons and regression analysis. This result is consistent with the results of Innes and Mitra (2012), as, in US-American sample, they found a strong contagion effect of dishonesty in a deception sender-receiver game (Gneezy, 2005)³⁸. This degree of contagion was present only for high levels of dishonesty (85%), higher than in the present work, which helps understand the low degree of contagion found here.

To look further into this result and explore the mechanism behind contagion, we used computational modelling and tested a total of 20 models categorized in four different types. This analysis revealed that, for most subjects, the best fitting model was one in which moral preferences, as indicated by a self-interest parameter and a moral cost parameter, were fixed for the whole task, with an additional contagion parameter that introduced a preference bias.

³⁸ This game involves two individuals: a sender and a receiver. The sender has private information on the payoffs of two or more options that differ in how lucrative they are to each participant. The sender then has to decide on a message to send to the receiver, who will in turn decide on one option. In its original formulation, the sender chooses between two messages, which inform the receiver on which option is the most lucrative for the receiver.

This parameter was null in the first block and was allowed to vary between the honest and dishonest block, being negative in the former and positive in the latter. We simulated the group behaviour using the Basic model, which encompasses a self-interest and a moral cost parameter, weighed by relative payoff for lying and absolute payoff for lying, respectively. This simulation was based on the results of the baseline block of an on-site pilot study. It should be noted that the best fitting model in the present sample effectively reduces to the Basic one when looking solely at baseline, thus converging with our previous results.

This result suggests a tendency for a lower subjective utility for being dishonest when facing the honest group and higher subjective utility when facing the dishonest group, analogous to what was found in Chung et al. (2015) in the domain of risk preferences. In their model, a constant of utility was added to one of the two options that the subject could choose from, weighing the individual towards or against it. This model was more suited to explain their results than a model that relied on blind following or a model in which risk preferences changed (i.e., where there was a change in the shape of the individual utility function, and not just a shift). Here, while for the majority of the subjects contagion acted in a simple manner, through a shift towards one of the options, for some few subjects the best fitting model was a more sophisticated RL model ($N = 4$; $\sim 6\%$). This further strengthens the notion that participants show heterogeneous behaviour. What is behind these individual differences is an open question.

In the present sample, the effect of contagion was small, suggesting that, in general, people may be less willing to trade their moral values and to be influenced by others, derived from the saliency of the character of the decision, as compared to other types of behaviour (e.g., Devaine & Daunizeau, 2017; Suzuki et al., 2016). Furthermore, there is no objective solution to a moral dilemma; as such, there is a lower informational effect to the social information conveyed to the participants, in the sense of Toelch and Dolan's (2015) classification. It should also be noted that it might have been obvious to some participants what the purpose of the experiment was, rendering them more attentive to their own moral values and more dedicated to protecting them, by standing their ground. Indeed, some post-experimental comments of participants on the on-site pilot study reveal that this might have happened, as around one-third of the participants indicated that they did not want to “sell their moral values”.

The present study was subjected to a number of limitations. First, there is a setback inherent in modelling, the best fitting model being considered the best from a set of tested ones. This offers no guarantees as to whether it is a good model, as there is an infinite

number of possibilities that could provide a better explanation for the results (Palminteri, Wyart, & Koechlin, 2017). Here, we tested a group of models that allowed us to answer a set of hypotheses, namely the nature of contagion (dynamic-RL or not) and the relationship to moral preferences and conformity. We attempted to overcome this limitation by falsifying our model via simulation, and using the subject-specific estimated parameters we were able to derive a similar average lying and behavioural tendencies. Still, this model had a low predictive accuracy, which suggests that there are important aspects of behaviour that are not being captured, perhaps resulting from the low degree of contagion we found. Rather than being a final solution to our present questions, this model is more appropriately seen as a step towards a comprehensive view of contagion in moral decisions.

We saw that individuals learned group behaviour easily, but that the prevalent norm in the honest group was easier to learn than the prevalent norm in the dishonest group. In truth, the norm of the dishonest group had greater variance than the norm of the honest group³⁹, which renders this result somewhat expected, and prompts a limitation in our paradigm. Since the amount of information received from both norms is not the same, an asymmetry in how individuals learn is created. This limits the interpretability of our, especially given that there might already be differences in how individuals update positive and negative moral information about others (Siegel, Mathys, Rutledge, & Crockett, 2018)⁴⁰. Still, it should be noted that in the second part of each block the norm was thoroughly learned.

Other limitations should be mentioned, namely the fact that we used the exact same task parameters in the Self and Predict trials, which might promote anchoring or availability effects in behaviour by allowing individuals to remember explicitly the decisions of others and to be influenced by them solely on the basis of having a reference for behaviour. We attempted to circumvent this by generating two sequences of trials in which the participant did not decide for oneself after seeing that combination of task parameters in the previous two Predict trials. Furthermore, decisions in Self trials were binary, while in Predict trials the information received was not binary. Still, to preclude this possibility entirely and strengthen the notion that individuals abstracted group behaviour, different task parameters could be used (as in Devaine & Daunizeau, 2017).

³⁹ Honest norm: $M = 9.200$, $SD = 0.913$ truth-tellers; Dishonest norm: $M = 2.200$, $SD = 3.227$ truth-tellers, in groups of 10 (Table A3 in Appendix A).

⁴⁰ The authors proposed an adaptive mechanism in which negative moral impressions (“beliefs about bad people”) are more uncertain and updated in a more flexible manner, perhaps to facilitate forgiveness and favourable updating, at least for mild moral transgressions.

Furthermore, the score on the comprehension questionnaire was not as high as desired, which indicates that some participants might not have understood the task. The fact that the data was collected online imposed a restraint on how much we could explain the task to the participants and answer their questions before they started the experiment. In particular, one participant, in a posterior exchange of e-mails, indicated that it was not clear to him that an error (i.e., lying) was paid. This sort of behaviour could confound our results, because it becomes less obvious whether subjects told the truth because of their honesty values or because they did not understand that they could lie, even if being confronted with others' decisions should have drawn attention to the fact that lying was an option (Fosgaard et al., 2013). On the other hand, individuals who did not consider the possibility of lying in this experimental setting might also not consider it in real-life situations, at least with (monetary) incentives of this magnitude.

Lastly, it should not be forgotten that the present data was collected during a pandemic and, for various reasons, this might affect how willing to lie subjects are. As, at the moment of writing, we are still in the midst of this crisis, it is difficult to ascertain the impact and long-lasting effects of the pandemic in the population (but see Chan-Chee et al., 2020; Fiorillo & Gorwood, 2020). For instance, it might be the case that some participants faced increasing economic difficulties due to lay-off response, as well as increased uncertainty about the future, which rendered them more likely to lie for money. Further, visceral and emotional states might play a role in how willing individuals are to engage in honest behaviour (Vincent, Emich, & Goncalo, 2013). Given that we did not control for mental illness, mental well-being, nor financial situation, hypotheses regarding these circumstances in our sample are impossible to exclude.

In general, the present study reveals a heterogeneous view of moral contagion, with a very restricted amount of others' influence over behaviour. Furthermore, most people were influenced in a very simplistic manner. Future studies should explore the individual characteristics that predict both honesty and conformity behaviour (for instance, it has been shown that social preferences might play a role; Maggian & Villeval, 2016) and both validate and falsify alternative models. The initial purpose for developing this paradigm was to explore and identify the neural networks associated with moral contagion and, in particular, disrupt contagion by means of tDCS. For this purpose, a localiser fMRI study would be run beforehand. At this point, we can only hypothesize which brain area would be targeted.

Given that the dlPFC is most likely involved both in honest and conformity behaviour (Greene & Paxton, 2009; Ruff et al., 2013), as it plays a wider and more general role in the exercise of cognitive control and deliberate thinking, targeting the dlPFC would most likely influence both processes. As such, the preferred target area would be the medial prefrontal cortex, which is involved in prosocial behaviour (Liao, Wu, Luo, Guan, & Cui, 2018) and has been implicated in detecting norm violation and behavioural alignment (Klucharev, Hytönen, Rijpkema, Smidts, & Fernández, 2009b; Klucharev et al., 2011; Wei et al., 2013; Wei, Zhao, & Zheng, 2017). In particular, the dorsomedial prefrontal cortex (dmPFC) has been implicated in conformity behaviour (Wu, Luo, & Feng, 2016), encoding the difference between a person's current preference and a cognitively balanced state, where the individual is aligned with the group (Izuma & Adolphs, 2013). As such, this area is more likely to be specifically involved in conformity. Nonetheless, the medial prefrontal cortex (mPFC) might be involved in regulating the sense of morality (Yuan et al., 2017).

As with any other complex cognitive process, moral decision-making should depend on a high number of factors, both environmental (social, cultural and economic) and biological (genetic and neurophysiological). Not surprisingly, literature on this topic is disjointed and relies on input from different fields, and it is still unclear how the results come together in a coherent framework. The present study and paradigm intend to contribute to this growing and interdisciplinary body of literature by providing a first attempt at characterising moral contagion in the honesty domain, resorting to computational modelling.

In summary, we employed an adapted die-roll task, with an additional norm manipulation, to uncover how behavioural contagion develops in the (dis)honesty domain. We found inter-individual heterogeneity in behaviour, but an overall low amount of lying. We saw that participants were fairly good at learning the prevalent social norm in each group – the honest and dishonest – and that they were slightly influenced by others' dishonest behaviour. Through computational modelling, we discovered that contagion was better accounted for by introducing a bias in preferences towards or against lying, context-independent, resulting from receiving information regarding the honest or the dishonest group's behaviour. This suggests that, in the moral domain, people may be less willing to conform than in other domains, a finding that is both comforting and frightening.

5. References

- Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, 87(4), 1115–1153. <https://doi.org/10.3982/ecta14673>
- Apps, M. A. J., & Ramnani, N. (2017). Contributions of the medial prefrontal cortex to social influence in economic decision-making. *Cerebral Cortex*, 27(9), 4635–4648. <https://doi.org/10.1093/cercor/bhx183>
- Aronson, E., Wilson, T. D., Akert, R. M., & Sommers, S. R. (2016). *Social Psychology* (9th ed.). Pearson Education.
- Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgement. In H. Guetzkow (Ed.), *Groups, leadership, and men* (pp. 76–92). Pittsburgh, PA: Carnegie Press.
- Asch, S. E. (1956). Studies of Independence and Conformity: I. A Minority of One Against a Unanimous Majority. *Psychological Monographs*, 70(9), Whole No. 416. <https://doi.org/10.1109/ICCE-China.2014.7029897>
- Bandura, A. (1977). *Social Learning Theory*. New Jersey: Prentice-Hall. <https://doi.org/10.16309/j.cnki.issn.1007-1776.2003.03.004>
- Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through Imitation of aggressive models. *Journal of Abnormal and Social Psychology*, 63(3), 575–582. <https://doi.org/10.1037/h0045925>
- Battigalli, P., Charness, G., & Dufwenberg, M. (2013). Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93, 227–232.
- Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature Letters*, 456(November), 245–250. <https://doi.org/10.1038/nature07538>
- Berns, G. S., Capra, C. M., Moore, S., & Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. *NeuroImage*, 49(3), 2687–2696. <https://doi.org/10.1016/j.neuroimage.2009.10.070>
- Berns, G. S., Chappelow, J., Zink, C. F., Pagnoni, G., Martin-Skurski, M. E., & Richards, J. (2005). Neurobiological correlates of social conformity and independence during mental rotation. *Biological Psychiatry*, 58(3), 245–253. <https://doi.org/10.1016/j.biopsych.2005.04.012>
- Bond, R. (2005). Group size and conformity. *Group Processes and Intergroup Relations*, 8(4), 331–354. <https://doi.org/10.1177/1368430205056464>

- Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the opinion of others affects our valuation of objects. *Current Biology*, 20(13), 1165–1170. <https://doi.org/10.1016/j.cub.2010.04.055>
- Chan-Chee, C., Léon, C., Lasbeur, L., Lecrique, J.-M., Raude, J., Arwidson, P., & Roscoat, E. du. (2020). The mental health of the French facing the COVID-19 crisis: prevalence, evolution and determinants of anxiety disorders during the first two weeks of lockdown. *Bulletin Épidémiologique Hebdomadaire*, 13ENG, 1–9.
- Chung, D., Christopoulos, G. I., King-Casas, B., Ball, S. B., & Chiu, P. H. (2015). Social signals of safety and risk confer utility and have asymmetric effects on observers' choices. *Nature Neuroscience*, 18(6), 912–916. <https://doi.org/10.1038/nn.4022>
- Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, 55(1), 591–621. <https://doi.org/10.1146/annurev.psych.55.090902.142015>
- Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. *Journal of Personality and Social Psychology*, 58(6), 1015–1026. <https://doi.org/10.1037/0022-3514.58.6.1015>
- Crockett, M. J. (2016). Computational Modeling of Moral Decisions. In *The Social Psychology of Morality* (pp. 71–90). Routledge.
- Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, 111(48), 17320–17325. <https://doi.org/10.1073/pnas.1408988111>
- Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, 20(6), 879–885. <https://doi.org/10.1038/nn.4557>
- Daunizeau, J., Friston, K. J., & Kiebel, S. J. (2009). Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D: Nonlinear Phenomena*, 238(21), 2089–2118. <https://doi.org/10.1016/j.physd.2009.08.002>
- Daunizeau, Jean. (2020). MBB-team/VBA-toolbox. GitHub. Retrieved from Jean Daunizeau (2020). MBB-team/VBA-toolbox (<https://github.com/MBB-team/VBA-toolbox>), GitHub. Retrieved July 11, 2020.
- Daunizeau, Jean, Adam, V., & Rigoux, L. (2014). VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. *PLoS Computational Biology*, 10(1). <https://doi.org/10.1371/journal.pcbi.1003441>

- Decety, J., & Yoder, K. J. (2017). The Emerging Social Neuroscience of Justice Motivation. *Trends in Cognitive Sciences*, 21(1), 6–14. <https://doi.org/10.1016/j.tics.2016.10.008>
- Devaine, M., & Daunizeau, J. (2017). Learning about and from others' prudence, impatience or laziness: The computational bases of attitude alignment. *PLoS Computational Biology*, 13(3), 1–28. <https://doi.org/10.1371/journal.pcbi.1005422>
- Dogan, A., Morishima, Y., Heise, F., Tanner, C., Gibson, R., Wagner, A. F., & Tobler, P. N. (2016). Prefrontal connections express individual differences in intrinsic resistance to trading off honesty values against economic benefits. *Scientific Reports*, 6(September), 1–12. <https://doi.org/10.1038/srep33263>
- Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4), 99–117. <https://doi.org/10.1007/978-1-349-62397-6>
- Erdfelder, E., Faul, F., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191.
- Fehr, E., & Fischbacher, U. (2003). Third Party Sanction and Social Norms. *Evolution and Human Behavior*, 25(2004), 63–87.
- Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, 8(4), 185–190. <https://doi.org/10.1016/j.tics.2004.02.007>
- FeldmanHall, O., & Mobbs, D. (2015). *A Neural Network for Moral Decision Making. Brain Mapping: An Encyclopedic Reference* (Vol. 3). Elsevier Inc. <https://doi.org/10.1016/B978-0-12-397025-1.00180-9>
- FeldmanHall, Oriel, Son, J.-Y., & Heffner, J. (2018). Norms and the Flexibility of Moral Action. *Personality Neuroscience*, 1(May). <https://doi.org/10.1017/pen.2018.13>
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.
- Fiorillo, A., & Gorwood, P. (2020). The consequences of the COVID-19 pandemic on mental health and implications for clinical practice. *European Psychiatry : The Journal of the Association of European Psychiatrists*, 63(1), e32. <https://doi.org/10.1192/j.eurpsy.2020.35>
- Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association*, 11(3), 525–547. <https://doi.org/10.1111/jeea.12014>

- Fischbacher, U., & Heusi, F. (2008). Lies in Disguise: An experimental study on cheating. *Research Paper Series*, 40, 1–19. Retrieved from <http://ideas.repec.org/p/twi/respas/0040.html>
- Foot, P. (2002). The Problem of Abortion and the Doctrine of the Double Effect. *Virtues and Vices and Other Essays in Moral Philosophy*.
- Fosgaard, T. R., Hansen, L. G., & Piovesan, M. (2013). Separating Will from Grace: An experiment on conformity and awareness in cheating. *Journal of Economic Behavior and Organization*, 93, 279–284. <https://doi.org/10.1016/j.jebo.2013.03.027>
- Fumagalli, M., & Priori, A. (2012). Functional and clinical neuroanatomy of morality. *Brain*, 135(7), 2006–2021. <https://doi.org/10.1093/brain/awr334>
- Garrett, N., Lazzaro, S. C., Ariely, D., & Sharot, T. (2016). The brain adapts to dishonesty. *Nature Neuroscience*, 19(12), 1727–1732. <https://doi.org/10.1038/nn.4426>
- Gary, B., Charness, G., & Dufwenberg, M. (2006). Promises & Partnership. *Econometrica*, 74(6), 1690–1601.
- Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The Truth About Lies: A Meta-Analysis on Dishonest Behavior. *Psychological Bulletin*, 145(1), 1–44.
- Gibson, R., Tanner, C., & Wagner, A. F. (2013). Preferences for Truthfulness: Heterogeneity among and within Individuals. *American Economic Review*, 103(1), 532–548. <https://doi.org/10.1257/aer.103.1.532>
- Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological Science*, 20(3), 393–398. <https://doi.org/10.1111/j.1467-9280.2009.02306.x>
- Gino, F., Gu, J., & Zhong, C. B. (2009). Contagion or restitution? When bad apples can motivate ethical behavior. *Journal of Experimental Social Psychology*, 45(6), 1299–1302. <https://doi.org/10.1016/j.jesp.2009.07.014>
- Gneezy, U. (2005). Deception: The Role of Consequences. *American Economic Association*, 95(1), 384–394.
- Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review*, 108(2), 419–453. <https://doi.org/10.1257/aer.20161553>
- Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences of the United States of America*, 106(30), 12506–12511. <https://doi.org/10.1073/pnas.0900152106>
- Heintz, C., Karabegovic, M., Molnar, A., & Heintz, C. (2016). The Co-evolution of Honesty

- and Strategic Vigilance. *Frontiers in Psychology*, 7(1503), 1–13.
<https://doi.org/10.3389/fpsyg.2016.01503>
- Hilbig, B. E., & Hessler, C. M. (2013). What lies beneath: How the distance between truth and lie drives dishonesty. *Journal of Experimental Social Psychology*, 49(2), 263–266.
<https://doi.org/10.1016/j.jesp.2012.11.010>
- Holm, S. (1979). Board of the Foundation of the Scandinavian Journal of Statistics A Simple Sequentially Rejective Multiple Test Procedure A Simple Sequentially Rejective Multiple Test Procedure. *Source: Scandinavian Journal of Statistics Scand J Statist*, 6(6), 65–70. Retrieved from
<http://www.jstor.org/stable/4615733>
<http://www.jstor.org/page/info/about/policies/terms.jsp>
<http://www.jstor.org>
- Innes, R., & Mitra, A. (2012). Is Dishonesty Contagious? *Economic Inquiry*, 51(1), 722–734.
<https://doi.org/10.1111/j.1465-7295.2012.00470.x>
- Insko, C. A., Smith, R. H., Alicke, M. D., Wade, J., & Taylor, S. (1985). Conformity and Group Size. *Personality and Social Psychology Bulletin*, 11(1), 41–50.
<https://doi.org/10.1177/0146167285111004>
- Izuma, K., & Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron*, 78(3), 563–573. <https://doi.org/10.1016/j.neuron.2013.03.023>
- Jacobsen, C., Fosgaard, T. R., & Pascual-Ezama, D. (2018). Why Do We Lie? a Practical Guide To the Dishonesty Literature. *Journal of Economic Surveys*, 32(2), 357–387.
<https://doi.org/10.1111/joes.12204>
- Kajackaite, A., & Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior*, 102, 433–444. <https://doi.org/10.1016/j.geb.2017.01.015>
- Kant, I. (2006). *Groundwork of the Metaphysics of Morals*. (M. Gregor, Ed.) (11th ed.). Cambridge: Cambridge University Press.
- Khalvati, K., Mirbagheri, S., Park, S. A., Dreher, J.-C., & Rao, R. P. N. (2019). A Bayesian Theory of Conformity in Collective Decision Making. In *Advances in Neural Information Processing Systems* (pp. 9699–9708).
- Khalvati, K., Park, S. A., Mirbagheri, S., Philippe, R., Sestito, M., Dreher, J.-C., & Rao, R. P. N. (2019). Modeling Other Minds: Bayesian Inference Explains Human Choices in Group Decision Making. *Science Advances*, *in press*(November), aax8783.
<https://doi.org/10.1101/419515>
- Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009a). Reinforcement Learning Signal Predicts Social Conformity. *Neuron*, 61(1), 140–151.

- <https://doi.org/10.1016/j.neuron.2008.11.027>
- Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009b). Reinforcement Learning Signal Predicts Social Conformity. *Neuron*, 61(1), 140–151. <https://doi.org/10.1016/j.neuron.2008.11.027>
- Klucharev, V., Munneke, M. A. M., Smidts, A., & Fernández, G. (2011). Downregulation of the posterior medial frontal cortex prevents social conformity. *The Journal of Neuroscience*, 31(33), 11934–11940. <https://doi.org/10.1523/JNEUROSCI.1869-11.2011>
- Kobis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D. G., & Shalvi, S. (2019). Intuitive Honesty Versus Dishonesty: Meta-Analytic Evidence. *Perspectives on Psychological Science*, 14(5), 778–796.
- Kohlberg, L. (1981). *The Philosophy of Moral Development: Moral Stages and the Idea of Justice. Essays on Moral Development* (Vol. II). San Francisco: Harper & Row, Publishers.
- Kohlberg, L. (1984). *The Psychology of Moral Development: The Nature and Validity of Moral Stages* (Vol. II). San Francisco: Harper & Row, Publishers. Retrieved from <https://www.worldcat.org/title/essays-on-moral-development-vol-2-the-psychology-of-moral-development-the-nature-and-validity-of-moral-stages/oclc/715523424>
- Krupka, E., & Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, 30(3), 307–320. <https://doi.org/10.1016/j.joep.2008.11.005>
- Kurtiness, W., & Pimm, J. B. (1983). The Moral Development Scale: A piagetian measure of moral judgment. *Educational and Psychological Measurement*, 43.
- Lelieveld, G. J., Shalvi, S., & Crone, E. A. (2016). Lies that feel honest: Dissociating between incentive and deviance processing when evaluating dishonesty. *Biological Psychology*, 117, 100–107. <https://doi.org/10.1016/j.biopsycho.2016.03.009>
- Leloup, L., Dongo Miletich, D., Andriet, G., Vandermeeren, Y., & Samson, D. (2016). Cathodal transcranial direct current stimulation on the right temporo-parietal junction modulates the use of mitigating circumstances during moral judgments. *Frontiers in Human Neuroscience*, 10(July). <https://doi.org/10.3389/fnhum.2016.00355>
- Liao, C., Wu, S., Luo, Y. jia, Guan, Q., & Cui, F. (2018). Transcranial direct current stimulation of the medial prefrontal cortex modulates the propensity to help in costly helping behavior. *Neuroscience Letters*, 674(June), 54–59. <https://doi.org/10.1016/j.neulet.2018.03.027>

- Lockwood, P. L., Klein-flügge, M., Abdurahman, A., & Crockett, M. J. (2019). Neural signatures of model-free learning when avoiding harm to self and other. *BioRxiv*, (718106). <https://doi.org/10.1101/718106>
- Luce, D. R., & Raiffa, H. (1957). *Games and Decisions: Introduction and critical survey*. New York: John Wiley & Sons, Inc.
- Maggian, V., & Villeval, M. C. (2016). Social preferences and lying aversion in children. *Experimental Economics*, 19(3), 663–685. <https://doi.org/10.1007/s10683-015-9459-7>
- Maréchal, M. A., Cohn, A., Ugazio, G., & Ruff, C. C. (2017). Increasing honesty in humans with noninvasive brain stimulation. *Proceedings of the National Academy of Sciences of the United States of America*, 114(17), 4360–4364. <https://doi.org/10.1073/pnas.1614912114>
- Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, 45(6), 633–644.
- Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin*, 34(7), 913–923. <https://doi.org/10.1177/0146167208316691>
- Obeso, I., Moisa, M., Ruff, C. C., & Dreher, J. C. (2018). A causal role for right temporoparietal junction in signaling moral conflict. *ELife*, 7, 1–16. <https://doi.org/10.7554/eLife.40671>
- Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, 21(6), 425–433. <https://doi.org/10.1016/j.tics.2017.03.011>
- Park, S. A., Goïame, S., O'Connor, D. A., & Dreher, J. C. (2017). Integration of individual and social information for decision-making in groups of different sizes. *PLoS Biology*, 15(6), 1–28. <https://doi.org/10.1371/journal.pbio.2001958>
- Park, S. A., Sestito, M., Boorman, E. D., & Dreher, J.-C. (2019). Neural computations underlying strategic social decision-making in groups. *Nature Communications*, 10(1), 1–12. <https://doi.org/10.1038/s41467-019-12937-5>
- Piaget, J. (1948). *The Moral Judgement of the Child*. Illinois: The Free Press.
- Qu, C., Hu, Y., Tang, Z., Derrington, E., & Dreher, J.-C. (2019). Neurocomputational mechanisms underlying immoral decisions benefiting self or others. *BioRxiv*, 832659. <https://doi.org/10.1101/832659>
- Qu, C., Météreau, E., Butera, L., Villeval, M. C., & Dreher, J. C. (2019). Neurocomputational mechanisms at play when weighing concerns for extrinsic rewards, moral values, and

- social image. *PLoS Biology*, 17(6), 1–27. <https://doi.org/10.1371/journal.pbio.3000283>
- Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, 45, 181–196. <https://doi.org/10.1016/j.joep.2014.10.002>
- Ross, D., & Brown, L. (2009). *Aristotle: Nicomachean ethics* (6th ed.). Oxford: Oxford University Press. <https://doi.org/10.1017/UPO9781844653584.004>
- Ruff, C. C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science*, 342(6157), 482–484. <https://doi.org/10.1126/science.1241399>
- Ruff, Christian C. (2018). Brain Stimulation Studies of Social Norm Compliance: Implications for Personality Disorders? *Psychopathology*, 51(2), 105–109. <https://doi.org/10.1159/000486898>
- Rush, A. J., First, M. B., & Blacker, D. (Eds.). (2008). *Handbook of Psychiatric Measures* (2nd ed.). American Psychiatric Pub.
- Saxe, R., & Powell, L. J. (2006). It's the Thought That Counts They Take the Prize. *Psychological Science*, 324(June), 2009.
- Sen, A. (2012). *Sobre Ética e Economia*. (E. Almedina, Ed.).
- Shalvi, S., Dana, J., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2), 181–190. <https://doi.org/10.1016/j.obhdp.2011.02.001>
- Sherif, M. (1936). The psychology of social norms.
- Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, 2(10), 750–756. <https://doi.org/10.1038/s41562-018-0425-1>
- Smith, A. (2009). *The Theory of Moral Sentiments*. Uplifting Publications. <https://doi.org/10.1002/9781118011690.ch10>
- Suzuki, S., Jensen, E. L. S., Bossaerts, P., & O'Doherty, J. P. (2016). Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. *Proceedings of the National Academy of Sciences of the United States of America*, 113(14), 3755–3760. <https://doi.org/10.1073/pnas.1600092113>
- Toelch, U., & Dolan, R. J. (2015). Informational and Normative Influences in Conformity from a Neurocomputational Perspective. *Trends in Cognitive Sciences*, 19(10), 579–589. <https://doi.org/10.1016/j.tics.2015.07.007>

- Vermunt, R. (2016). *The good, the bad, and the just: How modern men shape their world*. Routledge.
- Vincent, L. C., Emich, K. J., & Goncalo, J. A. (2013). Stretching the Moral Gray Zone: Positive Affect, Moral Disengagement, and Dishonesty. *Psychological Science*. <https://doi.org/10.1177/0956797612458806>
- Volz, L. J., Welborn, B. L., Gobel, M. S., Gazzaniga, M. S., & Grafton, S. T. (2017). Harm to self outweighs benefit to others in moral decision making. *Proceedings of the National Academy of Sciences of the United States of America*, 114(30), 7963–7968. <https://doi.org/10.1073/pnas.1706693114>
- Wei, Z., Zhao, Z., & Zheng, Y. (2013). Neural mechanisms underlying social conformity in an ultimatum game. *Frontiers in Computational Neuroscience*, 7(December), 1–7. <https://doi.org/10.3389/fnhum.2013.00896>
- Wei, Z., Zhao, Z., & Zheng, Y. (2017). The Neural Basis of Social Influence in a Dictator Decision. *Frontiers in Psychology*, 8(2134), 1–13. <https://doi.org/10.3389/fpsyg.2017.02134>
- Wheatley, T., & Decety, J. (2015). *The Moral Brain: A multidisciplinary perspective*. MIT Press.
- Wu, H., Luo, Y., & Feng, C. (2016). Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, 71, 101–111. <https://doi.org/10.1016/j.neubiorev.2016.08.038>
- Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America*, 107(15), 6753–6758. <https://doi.org/10.1073/pnas.0914826107>
- Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America*, 104(20), 8235–8240. <https://doi.org/10.1073/pnas.0701408104>
- Yuan, H., Tabarak, S., Su, W., Liu, Y., Yu, J., & Lei, X. (2017). Transcranial direct current stimulation of the medial prefrontal cortex affects judgments of moral violations. *Frontiers in Psychology*, 8(1812). <https://doi.org/10.3389/fpsyg.2017.01812>
- Zhu, L., Jenkins, A. C., Set, E., Scabini, D., Knight, R. T., Chiu, P. H., ... Hsu, M. (2014).

Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nature Neuroscience*, 17(10), 1319–1321. <https://doi.org/10.1038/nn.3798>

Zinchenko, O., & Arsalidou, M. (2018). Brain responses to social norms: Meta-analyses of fMRI studies. *Human Brain Mapping*, 39(2), 955–970.
<https://doi.org/10.1002/hbm.23895>

Appendix A – Additional Figures and Tables

Table A1

Marginal effects of Random-Effects Logistic Regression analysis on probability of lying, with subject as group variable

Predictor	Marginal effect
New cases	0.000 (0.000)
Total cases	0.000 (0.000)
New deaths	0.000 (0.000)
Total deaths	0.000* (0.000)

Note. Standard error, clustered at the subject level, is reported in parentheses (*** p < 0.01;

** p < 0.05; *p < 0.1). $\chi^2 (4) = 3.55$, prob > $\chi^2 = 0.470$.

Table A2

Proportion of males and females distributed per sequence

Sex	Sequence		
	1	2	Total
Male	15	16	31
Female	26	25	51
Total	41	41	82

Table A3

Number of people who reported truthfully in each group, according to task parameters

Relative gain from lying		Dice combination				
2€	6:1	1:2	2:3	3:4	4:5	
Honest Group	9	10	10	9	10	
Dishonest Group	8	8	9	9	6	
4€	6:2	1:3	2:4	3:5	1:3	
Honest Group	10	10	10	10	9	
Dishonest Group	2	4	2	2	4	
6€	6:3	1:4	2:5	6:3	1:4	
Honest Group	9	9	10	9	9	
Dishonest Group	0	0	0	0	1	
6€	6:4	1:5	6:4	1:5	6:4	
Honest Group	9	7	8	8	9	
Dishonest Group	0	0	0	0	0	
10€	6:5	6:5	6:5	6:5	6:5	
Honest Group	9	10	7	10	9	
Dishonest Group	0	0	0	0	0	

Note. Number of the left refers to true draw and on the right to the false draw. In bold are repetitions of parameters, to accomplish 5 trials per relative gain from lying. Honest norm: 9.200 ± 0.913 truth-tellers; Dishonest norm: 2.200 ± 3.227 truth-tellers.

Table A4

Results of Random-Effects Logistic Regression analysis on probability of lying, with subject as group variable

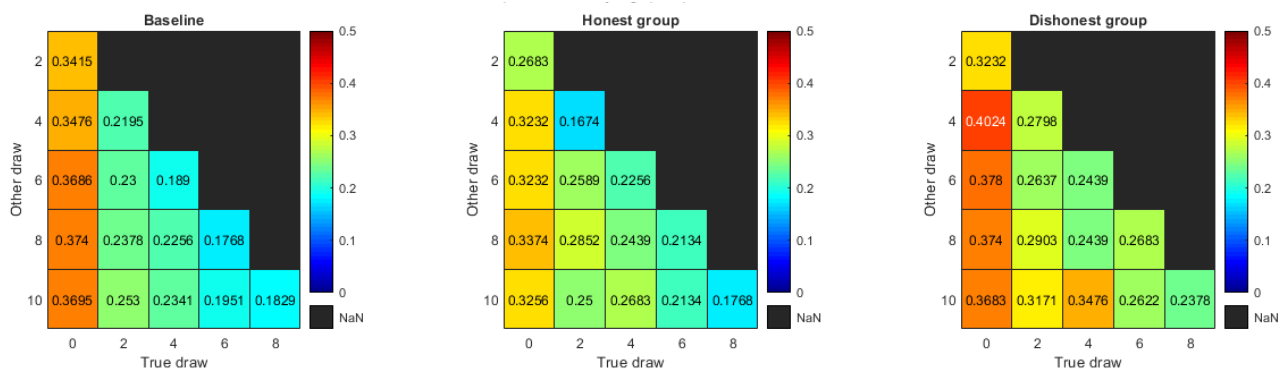
Predictor	Odds Ratio
Intercept	0.135 (0.392)
Dishonest Block First	5.876** (4.827)
Honest Block	0.805 (0.194)
Dishonest Block	1.574* (0.403)
True Draw	0.413*** (0.100)
False Draw	1.039 (0.031)
True Draw \times False Draw	1.061*** (0.021)

Note. Standard error, clustered at the subject level, is reported in parentheses (***) $p < 0.01$;

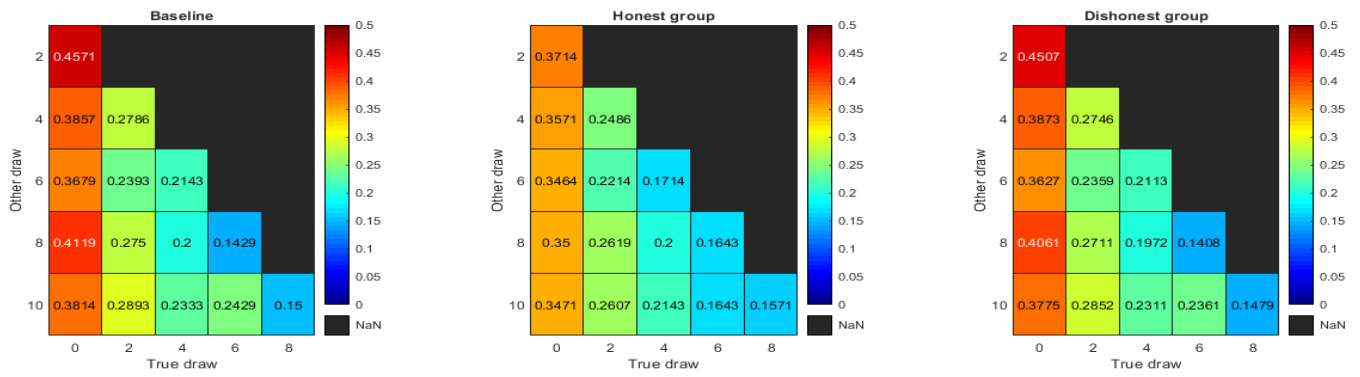
** $p < 0.05$; * $p < 0.1$).

Figure A1

Amount of lying per task parameters, per block



Note. The heatmaps suggests a difference in amount of lying in different blocks.

Figure A2*Amount of lying per task parameters, per block, as simulated*

Note. The heatmap suggests a difference in amount of lying between blocks, especially between the honest block and the other two.

Appendix B

Both instructions and comprehension questionnaire were originally in French. Here, we present a freely translated version.

1. Instructions

We thank you for participating in this online experiment about decision-making. Please try to complete the experiment at once with the least amount of pauses possible. In this experiment, all your decisions are anonymous. Your choices will not be associated to your name or other information that trace back to you. **Contrarily to most experiments, your decisions will determine your gains.** Apart from these gains, you will also receive a fixed amount of 1€. Other participants will not know how much you earned.

The experiment consists in a sequence of draws of a 6-faced die. What you should do afterwards depends on the type of trial you are in. There are two types of trials: Solo trials and Prediction trials. We will explain them successively.

Solo trials

In each Solo trial you will see the result of a draw of a 6-faced die for one second. You should memorise the random number because we will ask you to report it, to determine your payment. In the report screen, two numbers will be shown on the left and on the right of the screen. If you think the number on the right is the one that was shown previously, you should click on it, on the right. Conversely, if you think the number on the left is the one that was shown previously, you should click on the left. Each Solo trial is signalled by the presence of an icon on the top of the screen. That icon will be shown in the end of the instructions.

As indicated previously, your payment on each Solo trial will depend on the number you report. **More precisely, the number you report multiplied by two corresponds to your payment in euros.** If you report the number 1, then you receive 2€, number 2, 4€, number 3, 6€, number 4, 8€ and number 5, 10€. However, if you report number 6, you will receive nothing (0€).

Prediction trials

The start of a Prediction trial is the same as a Solo trial. You will see the result of a 6-faced die draw for one second. However, instead of reporting the result of this draw, you will **predict what a group of 10 people reported**. These people have made the same decisions as you in the Solo trials. They were not informed of others' decisions.

After seeing the draw, you will observe the screen that the group of 10 people has seen. This is the two numbers that were shown on screen. After you have estimated **the number of people that have correctly reported the result of the draw**, you should write this number in the dialogue box on the screen. To validate your estimate, press Enter. After your response is validated, you will be informed on the correct number of people that has correctly reported the result of the draw. If your prediction is incorrect, you will receive 0 euros. If your prediction is correct, you will receive 1 euro. Each Prediction trial is signalled by the presence of an icon on the top of the screen. That icon will be shown in the end of the instructions.

Structure of the session

The experiment is composed of three parts:

You will start by performing a few trials to familiarize yourself with the task. The decisions you make during this training phase will not be taken into account, neither for analyses, nor for your payment. **Furthermore, the parameters for these trials are fictitious.**

Afterwards, the experiment begins, and it consists of three blocks:

- 1) The first block is solely composed of Solo trials. It is divided in two parts of 25 trials each.
- 2) The second part is composed of Solo and Prediction trials: a first part of 25 Prediction and 25 Solo trials that are presented alternately. You will start by a Prediction trial, then a Solo trial, followed by a Prediction trial, and so on. After, a second part similar to the first.
- 3) The third and last block is organized in the same way as the second block.

Group composition

As previously mentioned, the second and third block are similar in their organization. Nonetheless, the group of 10 people whose decisions you will predict is not the same in the

two blocks. In the second block you will observe and predict the decisions of Group 1 and in the third block you will observe and predict the decisions of Group 2.

The people that form these two groups **have performed the same task as you in a previous study**. They have only performed block 1 of this experiment. They have not received information about others' decisions.

We have selected and separated them into two groups due to their behaviour. Each individual behaved in a way similar to the other group members. They were mostly recruited via the **Facebook page of the Institut des Sciences Cognitives**. They are mostly **students** and have ages between **18 and 35** years old.

To protect their identity, we cannot give you more information about these two groups of 10 individuals. In general, you can consider that **they are not different from you**.

Payment

A Solo trial and a Prediction trial will be randomly selected in the end of the three blocks. For the Solo trial, you will be paid according to the number you have reported. As for the Prediction trial, if you have predicted correctly, you will receive 2 euros, in addition to your potential gains for the Solo trial; if not, you will not receive an additional gain. Lastly, you will receive 1 euro for your participation in this experiment, independent of your decisions.

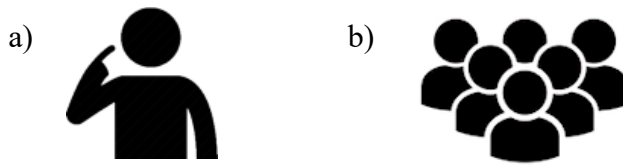
Other information

Remember: since you do not know which trials will be randomly selected, your best strategy is to treat each decision as if it was indeed the one which counts. Every decision is definite and cannot be changed.

Lastly, the people who are organizing this experiment will not know your secret decisions. Further, since your decisions are secret, your answers will not be controlled. They will not be used as Prediction trials for other participants.

Figure B1

Icons utilised for signaling trial type



Note. Icons presented on screen indicating that the participant is currently in a a) Self trial and b) Predict trial.

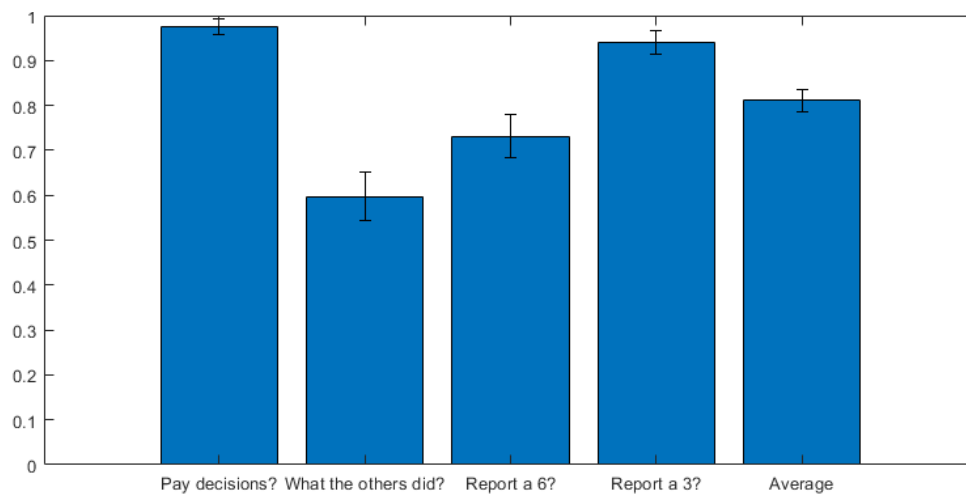
2. Comprehension questionnaire

- 1) Your payment depends on your decisions?
 - a. Yes
 - b. No
- 2) The people whose decisions you are going to observe have made:
 - a. Only Solo decisions
 - b. Only Predict decisions
 - c. Both
- 3) If you report a 6, you receive:
 - a. 0€
 - b. 6€
 - c. 12€
- 4) If you report a 3, you receive:
 - a. 0€
 - b. 6€
 - c. 12€

As can be seen in Figure B2, most participants understood that they will be paid according to their decisions (mean score of 0.976). On the other hand, they had more trouble understanding that these other participants made their decisions alone, i.e., where not exposed to other individuals' behaviour (mean score of 0.598).

Figure B2

Average score of each question of the comprehension questionnaire



Note. Questions are scored 1 if answered correctly and 0 if answered incorrectly. Overall mean score: 0.811; Question 1: 0.976; Question 2: 0.598; Question 3: 0.731; Question 4: 0.939.

Appendix C – Details on modelling

For all models, the following softmax decision rule applies on the probability of participant i lying on trial t :

$$P_i(\text{lie}) = \frac{1}{1 + e^{-\beta \Delta U_{i,t}}} \quad (1)$$

Table C1

Models tested on VBA

Class	Type of model	
	Standard	Multiblock
Static	Basic (1)	Basic (2)
Static		Basic- α (3)
Static		Basic- δ (4)
Static	Fixed Cost (5)	Fixed Cost (6)
Static		Fixed Cost- α (7)
Static		Fixed Cost- δ (8)
Static	Conformity (9)	Conformity (10)
Reinforcement Learning	Action- α (11)	Action- α (12)
Reinforcement Learning	Action- δ (13)	Action- δ (14)
Reinforcement Learning	Action- α - δ (15)	Action- α - δ (16)
Reinforcement Learning	Conformity Action (17)	Conformity Action (18)
Reinforcement Learning	Conformity Outcome (19)	Conformity Outcome (20)

1. Static Models

1.1. Basic Model

$$\Delta U_{i,t} = \alpha_i(\pi_{D,t} - \pi_{H,t}) - \delta_i \pi_{D,t} \quad (2)$$

Where $\pi_{D,t}$ and $\pi_{H,t}$ refer to payoff at trial t for being dishonest and honest, respectively. Furthermore, α is constrained to be a positive parameter, as it reflects self-interest, and δ is expected to be negative. In the multiblock version, both α and δ are estimated per block. In $-\alpha$ multiblock version, α is estimated per block while δ is fixed; and vice versa for the $-\delta$ multiblock version.

1.2. Fixed Cost Model

$$\Delta U_{i,t} = \alpha_i(\pi_{D,t} - \pi_{H,t}) - \delta_i \quad (3)$$

Where α is constrained to be a positive parameter, as it reflects self-interest, and δ is expected to be negative. In the multiblock version, both α and δ are estimated per block. In the $-\alpha$ multiblock version, α is estimated per block while δ is fixed; and vice versa for the $-\delta$ multiblock version. This model differs from the Basic Model because the moral cost δ parameter does not weight task parameters.

1.3. Conformity Model

$$\Delta U_{i,t} = \alpha_i(\pi_{D,t} - \pi_{H,t}) - \delta_i \pi_{D,t} + \gamma_i \quad (4)$$

Where α is constrained to be a positive parameter, as it reflects self-interest, and δ is expected to be negative. In this model, γ , a conformity parameter, is set to zero for block 1 (baseline), since in this block participants are not yet receiving information regarding others' decisions. In the multiblock version, α and δ are fixed and only γ is allowed to vary between the honest and dishonest block.

2. Reinforcement Learning Models

2.1. Moral- α Model

$$\Delta U_{i,t} = (\alpha_i + \alpha'_{i,t})(\pi_{D,t} - \pi_{H,t}) - \delta_i \pi_{D,t} \quad (5)$$

$$\alpha'_{i,t} = \alpha_{i,t-1} + \lambda(n_{\text{honest}} - n_{\text{dishonest}}) \quad (6)$$

As before, α is constrained to be a positive parameter, as it reflects self-interest, and δ is expected to be negative. In the multiblock version of this model, the learning rate λ is assumed to be different in the honest and dishonest block. Hidden states (i.e., $\alpha'_{i,t}$) are also estimated separately for each block. Note that as the number of honest people, relative to dishonest, increases (in particular, if there are more dishonest than honest people), the hidden state increases, thus resulting in a higher weight for the relative gain from lying in the utility function.

2.2. Moral- δ Model

$$\Delta U_{i,t} = \alpha_i(\pi_{D,t} - \pi_{H,t}) - (\delta_i + \delta'_{i,t})\pi_{D,t} \quad (7)$$

$$\delta'_{i,t} = \delta_{i,t-1} + \lambda(n_{\text{honest}} - n_{\text{dishonest}}) \quad (8)$$

Again, α is constrained to be a positive parameter, as it reflects self-interest, and δ is expected to be negative. In the multiblock version of this model, the learning rate λ is assumed to be different between the honest and dishonest block. Hidden states (i.e., $\delta'_{i,t}$) are also estimated separately for each block. Note that as the number of honest people, relative to dishonest, decreases (in particular, if there are more honest people than dishonest), so the hidden state decreases, thus resulting in a more negative weight of the influence of the absolute gain from lying in the utility function.

2.3. Moral- α - δ

$$\Delta U_{i,t} = (\alpha_i + \alpha'_{i,t})(\pi_{D,t} - \pi_{H,t}) - (\delta_i + \delta'_{i,t})\pi_{D,t} \quad (9)$$

$$\alpha'_{i,t} = \alpha_{i,t-1} + \gamma(n_{\text{honest}} - n_{\text{dishonest}}) \quad (10)$$

$$\delta'_{i,t} = \delta_{i,t-1} + \gamma(n_{\text{honest}} - n_{\text{dishonest}}) \quad (11)$$

Here, both effects described above happen simultaneously, that is, both self-interest α parameter and moral cost δ parameter are updated as a function of how many people decided to lie in the Predict trial. If more people decided to lie, relative to telling the truth, the self-interest parameter increases and the moral cost parameter decreases (i.e., its absolute value, as this parameter is expected to be negative); whereas the opposite pattern results in a decrease in self-interest and increase in moral cost. If exactly five people lied and were honest, no updating occurs.

2.4. Conformity-Action

$$\Delta U_{i,t} = \alpha_i(\pi_{D,t} - \pi_{H,t}) - \delta_i\pi_{D,t} + \gamma_i \times x'_{i,t} \quad (12)$$

$$x'_{i,t} = x_{i,t-1} + \lambda(n_{\text{honest}} - n_{\text{dishonest}}) \quad (13)$$

Where α is constrained to be a positive parameter, as it reflects self-interest, and δ is expected to be negative. Furthermore, γ is a conformity parameter, that weights $x'_{i,t}$, which is updated in a RL-fashion according to a learning rate parameter, λ . As in the static version, γ is set to zero for block 1 (baseline), since in this block participants are not yet receiving information regarding other's decisions. In the multiblock version of this model, learning rate, hidden states (i.e., $x'_{i,t}$), and the conformity parameter are estimated separately for the honest and the dishonest block.

2.5. Conformity-Outcome

$$\Delta U_{i,t} = \alpha_i(\pi_{D,t} - \pi_{H,t}) - \delta_i \pi_{D,t} + \gamma_i \times x'_{i,t} \quad (14)$$

$$x'_{i,t} = x_{i,t-1} + \lambda(n_{honest} - n_{dishonest})(\pi_{L,t} - \pi_{\bar{L},t}) \quad (15)$$

Where α is constrained to be a positive parameter, as it reflects self-interest, and δ is expected to be negative. In this version of the utility function, $x'_{i,t}$ is updated as a function of an interaction between the amount of people who decided to lie and the total payoff for lying (i.e., difference in payoff between lying and being honest). As in the static version, γ is set to zero for block 1 (baseline), since in this block participants are not yet receiving information regarding others' decisions. Note that $(\pi_{D,t} - \pi_{H,t})$ is always positive, as being dishonest is always more profitable than being honest. For instance, if more people were honest than dishonest, the hidden state will be updated more negatively (i.e., tendency to conform to honesty) as the payoff for lying increases.